nature

www.nature.com/nature

Vol 435|Issue no. 7040|19 May 2005

Running out of juice

Despite the hype, there's no sign that the Congress will produce an energy bill worthy of the formidable energy-policy challenges faced by the United States.

Suppose that one was charged with the task of framing US energy legislation from scratch, without the political baggage that has accumulated over several years' unsuccessful efforts to push such a law through the Congress.

One top priority, endorsed at both ends of the political spectrum, would undoubtedly be to reduce the United States' large and growing reliance on imported oil. Another — which unfortunately benefits from no such political consensus — would be to stabilize and then to reduce the nation's greenhouse-gas emissions. A third priority would be to support a balanced programme of energy-related research and development at a level commensurate with the massive scale of the US energy business, not to mention its long-term strategic importance.

A range of options for tackling these pressing concerns were spelt out robustly last year in the final report of the independently funded, bipartisan National Commission on Energy Policy (see www.energycommission.org).

But judging from the version of the bill passed by the House of Representatives last month, none of these concerns are being properly addressed by the Congress. For a start, the bill obstinately declines to take seriously the need to cut carbon emissions. It does nothing to contain the demand for oil through the most obvious mechanism, the imposition of modest fuel-efficiency standards for cars and light trucks. And it fails to authorize any substantial revival in energy research spending, which has been steadily run down from its peak in the aftermath of the 1974 energy crisis.

"The bill obstinately declines to take seriously the need to cut carbon emissions."

The Senate is this week releasing a measure that will purport to look at the bigger picture. A bipartisan bill being put forward jointly by Senator Pete Domenici (Republican, New Mexico) and Jeff Bingaman (Democrat, New Mexico) will take a more broadly based approach than the House bill, and boost spending on some aspects of energy research.

But even the Senate seems to lack the will to seriously confront the consequences of the US addiction to oil imports. The bill will provide generous subsidies for clean coal and nuclear-power generation, and less generous ones for renewable energy sources. But in a country where about two-thirds of energy use is for transportation, it seems unlikely to bite the bullet and take steps to reduce oil consumption. The Senate could start to address this by, for example, promoting new cellulose-based biofuels (not to be confused with the existing ethanol programme, which is designed primarily as a subsidy machine for corn farmers). It could also yet decide to insist on new fuel-efficiency standards.

But many leaders in the Congress seem to prefer to squabble over oil drilling in the Arctic National Wildlife Refuge in Alaska, rather than address the issue of oil consumption. The impression continues to grow that the president and congressional leaders want to win on the Alaskan refuge primarily to rile environmentalists, rather than to obtain oil. Oil companies are barely registering an interest, the political and economic costs being so high.

As is usual for legislation of this type, the House bill is laden with goodies for special interests, including some $2 billion for deep gas drilling, subsidies for those keen to build nuclear power stations, and various tax breaks. But even the corporate giveaways — which are valued by some critics at $8 billion over ten years — are paltry by Washington's standards of profligacy. This is further proof, if any were needed, that political leaders just haven't grasped the scale of the problem. ■

Iran's long march

Scientific excellence can re-emerge in Iran, unless there is political upheaval or further sanctions.

Stories about people who prevail against the odds are always heart-warming, and the tale in this issue of *Nature* about a group of Iranian neuroscientists (see page 264) is no exception. Such successes have been rare in Iran's recent history. Over the past few years, however, the country has been investing more public money in science, and nurturing the gradual emergence of a climate in which high-quality research can flourish.

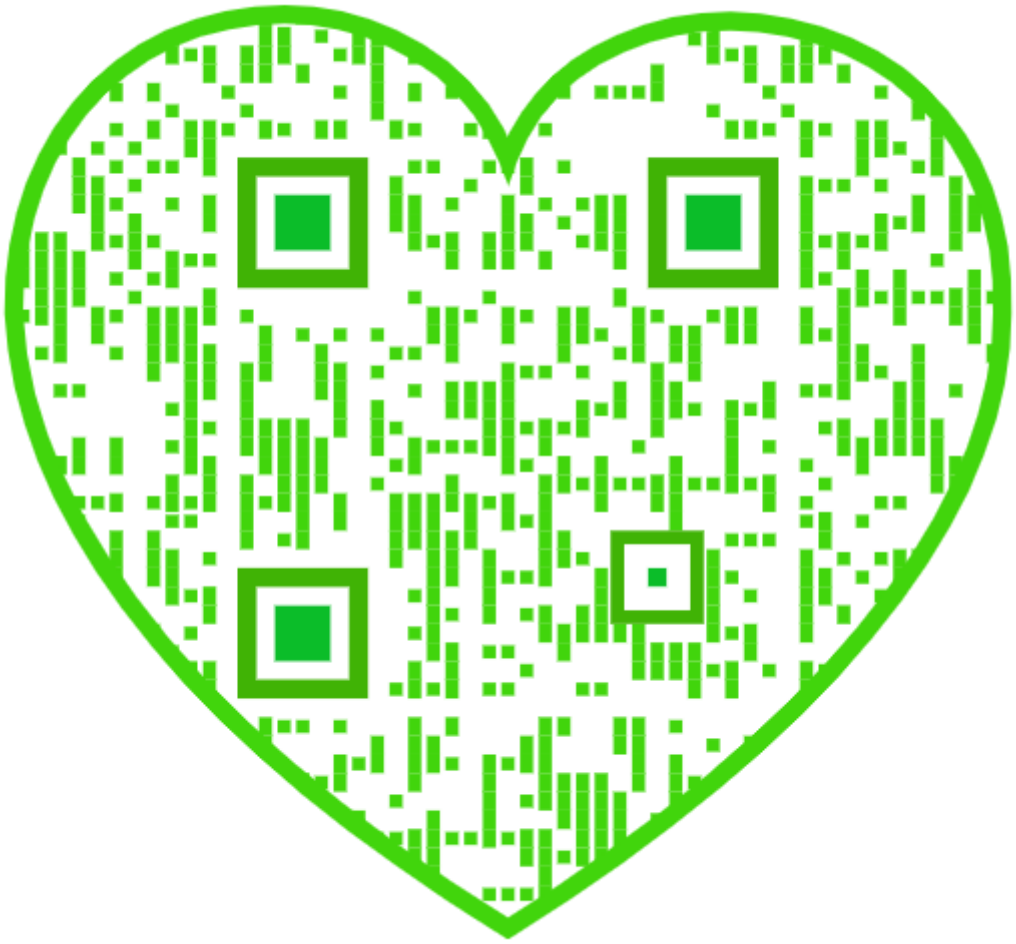
During this period, social and political constraints on Iranian life have visibly relaxed. Headscarves on the streets of Tehran are now sometimes perched so far back that Velcro is needed to stop them falling off altogether. Science has also benefited from this relaxation.

Even Ayatollah Ali Khamenei, the senior cleric effectively in charge of the country, has been speaking up for science, calling on his country to develop "self-confidence in all scientific fields". The call resonates within a small academic community that is keenly aware of Iran's rich and ancient heritage of scientific achievement.

The government, led by the elected president, Mohammad Khatami, has meanwhile increased science funding. It has also introduced reforms — approved this year, after a tough fight — that give universities and research institutes much more autonomy. In February, academics' salaries were sharply increased: as a result, many of them will be able to make ends meet without a second job, for the first time since the Islamic revolution in 1979.

Some of the most productive laboratories, such as the Institute for

©2005 Nature Publishing Group



Studies in Theoretical Physics and Mathematics (IPM) in Tehran and the Institute for Advanced Studies in Basic Sciences (IASBS) in Zanjan, have taken the opportunity presented by the reform package to reorganize themselves. The IPM says it is basing its new structure on that of the Max Planck Society in Germany, in which the research directors of individual labs have full control over budgets and research agendas.

The publication rate of Iranian scientists in international journals, meanwhile, has quadrupled over the past decade. Although still relatively low, the average impact factor of Iranian papers has also risen. But much of the improvement can be traced back to a small number of researchers. The number of papers per researcher at the IASBS, for example, is nearly twice that of its closest competitor, the Sharif University of Technology in Tehran. The reforms set out to extend the base of scientific excellence in Iran.

That may prove difficult to accomplish, given the current isolation of Iranian scientists from their peers abroad. US economic sanctions make it difficult for them to travel to meetings in the United States. And because most scientific equipment contains US components that are covered by sanctions, Iranians can only procure it through middle-men who disguise the equipment's ultimate destination but inflate its price and leave recipients with no quality guarantees or maintenance arrangements. Scientists in Iran tell stories of waiting up to two years for an order, only to receive shipment of the wrong model.

Unfortunately, given current tensions over Iran's nuclear programme (see *Nature* **435,** 132; 2005), the sanctions are unlikely to be lifted in the short term. Instead, the situation could rapidly deteriorate. Internal support for scientific development and for the free exchange of ideas is far from assured.

The president will be replaced after elections due on 17 June. One candidate for the presidency is immunologist Mostafa Moin, a former research minister who resigned from the government in 2003 in protest at parliamentary opposition to the university reform act. Moin could be relied on to continue along the reformist path set by Khatami. It is just as likely, however, that the elections will yield instead a president less interested in either reform or scientific freedom.

The Iranian scientists who have put together strong research groups have done so by carefully steering clear of politics. But they are admirably determined to help build up a research infrastructure in the country that will outlast their own careers. Scientists everywhere should avail themselves of every opportunity to assist these grassroots efforts. ■

Policing integrity

Plagiarism allegations should serve as reminders that universities cannot police misconduct on their own.

Serious scientific misconduct is a pernicious and under-reported problem that the research community has sought to police, with mixed success. There is little evidence, fortunately, that the falsification or fabrication of research data is becoming more frequent. However, there are rumblings that one type of misconduct is on the increase: plagiarism, made easier by the rapid electronic dissemination of research results (see page 258).

In many countries, the investigation of plagiarism and other misconduct remains fraught with difficulty. Universities are terrified of the adverse publicity that will accompany any finding of misconduct. And if the accused doesn't accept the outcome, they have the option of taking legal action against the accusers, the investigators or the institution in question.

In those circumstances, it is inevitable — although still unfortunate — that some senior university officials will opt for the easiest path, rather than the most righteous one. In some cases, an errant academic may be encouraged to leave an institution without full public disclosure of the events leading to the departure — a lamentable outcome that will simply serve to export misconduct, rather than to deter it.

In the United States, some defences are in place against this kind of thing. Biomedical research is overseen by the Office of Research Integrity (ORI) at the health department; research supported by the National Science Foundation, which funds most non-biomedical university research, is overseen by the Inspector General's office at the foundation. The ORI, in particular, has been active in helping universities run their own investigations properly. And if internal university probes fall short, these bodies can step in and carry out the investigation themselves.

The system established in the United States has never been quite as rigorous or effective as its architects had hoped. Yet the existence of these oversight organizations still puts the United States in a better position to police misconduct than most other nations.

In Britain, the universities themselves, through a group called Universities UK, are developing plans to establish a central office to oversee biomedical research. But the plans have already run into criticism for failing to let the body pursue its own investigations, and for the odd suggestion by the universities that the office accept sponsorship from the pharmaceutical industry (see *Nature* **434,** 263; 2005).

"It is inevitable — although still unfortunate — that some senior university officials will opt for the easiest path, rather than the most righteous one."

The major French research agencies have committees for dealing with misconduct, but none of them are as visible, well staffed or effective as their US counterparts, and Italy has no such arrangement at all. Germany has three full-time ombudsmen on hand to supervise university investigations. The German system was put in place after a prominent case of data fabrication in 1997, when it emerged that dozens of studies by Friedhelm Herrmann and Marion Brach, cancer researchers at the Max Delbrück Centre for Molecular Medicine in Berlin, contained falsified data.

It shouldn't take a recurrence of something like the Herrmann and Brach case for other leading scientific nations to ensure that they have adequate measures in place to investigate misconduct, properly publicize the findings, and punish the culprits. ■

©2005 Nature Publishing Group

RESEARCH HIGHLIGHTS

Fish scarper up north

Science doi:10.1126/science.1111322 (2005)

Climate change is forcing fish in the North Sea to shift to cooler waters, concludes a team of researchers led by Allison Perry and John Reynolds of the University of East Anglia, UK. The group looked at how the distribution of 36 bottom-dwelling species has changed over the past 25 years. Nearly two-thirds of them — including commercially exploited fish such as the Atlantic cod (*Gadus morhua*, pictured) and species not targeted by fisheries, such as the scaldfish (*Arnogossus laterna*) — have moved north or sunk to deeper waters.

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

OXFORD SCIENTIFIC

COMPUTER SIMULATION

Monte Carlo races

Phys. Rev. Lett. **94,** 170201 (2005)

Researchers don't like to admit defeat, but physicists grappling with a certain class of Monte Carlo computer simulations should be told that there's no easy way to speed them up.

Monte Carlo simulations are often used to look for equilibrium states in systems with many interacting components. But when those components are fermionic particles — electrons, for example — the time taken to find reliable solutions increases exponentially with the number of particles. This hampers simulations for calculating electronic structures, especially for systems (such as high-temperature superconductors) in which the electrons interact strongly.

Now, Swiss physicists Matthias Troyer and Uwe-Jens Wiese have proved that searches for a faster Monte Carlo method for fermions are doomed. They show that the problem belongs to the class of 'nondeterministic-polynomial' hard computational problems, for which no general short cuts exist.

MATERIALS

Stiff stuff

Nature Mater. doi:10.1038/nmat1392 (2005)

Cells resting on stiff surfaces are better at taking up genes delivered by non-viral carriers; they also express these genes more readily.

David Mooney's team at Harvard University delivered DNA strands to mouse cells using polyethyleneimine as a vector, monitoring the gene's delivery by a change in fluorescence as it detached from the vector. The cells were attached to various hydrogels, whose rigidity could be varied easily.

Although gene expression was not affected by the gels' surface chemistry, it increased markedly for more rigid gels. Exploiting this effect might boost the rate of gene transfection and expression from non-viral vectors.

CELL BIOLOGY

Achilles' heel for HIV

Immunity **22,** 607–619 (2005)

RNA silencing — already used by plants and invertebrates to fend off viral infection — is a tool proposed for treating HIV-1 infection in humans. At first, short interfering RNAs can inhibit HIV-1 replication in cell culture. But over time this inhibitory effect wanes, probably because HIV-1 mutates so rapidly. In new work, however, Kuan-Teh Jeang and colleagues, at the National Institute of Allergy and Infectious Diseases, identify an HIV-1 sequence that is protected from short interfering RNA attacks by a protein. This suggests that it cannot be altered for functional reasons. If researchers can find a way to evade the protector, the sequence may be a viable target for RNA interference therapy.

BIOLOGY

Stress ploys

PLoS Biol. doi: 10.1371/journal.pbio.0030176 (2005)

Bacteria induce mutations in their genome when put under stress. This finding might offer a way to combat the emergence of antibiotic resistance.

Research led by Floyd Romesberg at the Scripps Research Institute in La Jolla, California, revealed that the SOS response triggered by some antibiotics actively induces mutations in the bacterial genome. When the team interfered with one of the enzymes involved in this process, *Escherichia coli* did not evolve resistance to important quinolone and rifamycin antibiotics.

BIOTECHNOLOGY

Souped-up vectors

Nature Biotechnol. doi:10.1038/nbt1094 (2005)

Plants can be useful factories for manufacturing foreign proteins. But permanently modified crops carry environmental risks, and transient infection with a transgene is too inefficient to be of industrial use. Or so it was thought.

Researchers from Icon Genetics in Germany have shown how to modify a viral vector carrying the appropriate transgene to make it more potent. They made the vector, derived from the tobacco mosaic virus, more like the plant RNA by adding non-coding segments called introns. And they deleted segments that they suspected the plant nucleus would find hard to process. This increased export of transgenic RNA from infected nuclei, cutting 1,000-fold the number of vectors needed to make a cell produce the protein. The technique worked in a variety of different plants, including petunias, cucumbers and spinach.

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

©2005 Nature Publishing Group

THE IMAGE BANK/GETTY IMAGES

CELL BIOLOGY

Pore relation

Genes Dev. **19,** 1188–1198 (2005)

A controversial theory suggesting that genes relocate to the periphery of the nucleus when they are transcribed is supported by studies in yeast cells. A team led by Pamela Silver at Harvard Medical School tracked the location of the many genes whose expression is enhanced when yeast cells are exposed to a mating pheromone. The researchers show that most of the activated genes cluster around nuclear pores — openings through which molecules pass between the nucleus and cytoplasm. They also use high-resolution mapping to investigate how the chromosome hosting the genes changes its spatial orientation to put the genes in place.

NANOTECHNOLOGY

Into the groove

Nano Lett. doi:1021/nl050405n (2005)

Nanoscale ice sculptures could provide templates for making transistors and miniature machines, suggests a team at Harvard University. The researchers used electron-beam lithography to carve grooves through a 20-nanometre-thick layer of ice on a silicon wafer, and then applied a coating of chromium. A rinse in isopropyl alcohol removed the icy template and its metal overlayer, leaving strips of chromium less than 20 nanometres wide.

Unlike the polymer mask materials used to pattern silicon today, ice doesn't need spinning or baking on to the substrate. And it can be removed by sublimation, eliminating the need for solvents, although this process leaves stray chromium flakes.

VVG/SPL

CANCER

Healing hijacked

Cell **121,** 335–348 (2005)

The body's wound-healing process is subverted by breast-cancer tumours to help them grow and build a network of blood vessels, according to a study led by Robert Weinberg of the Whitehead Institute in Cambridge, Massachusetts.

Tumours are already known to contain fibroblasts — cells usually found only in inflamed or regenerating tissue that help build connective tissue. The team found that in invasive breast tumours, these fibroblasts make a protein called SDF-1 that encourages cancerous cells (pictured left) to grow. The protein also summons cells to generate the blood vessels that tumours need.

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

HYDRODYNAMICS

Skipping stones

Phys. Rev. Lett. **94,** 174501 (2005)

Stones, as many time-wasters know, skim across water when thrown the right way. But how to get the most bounces? Experiments reported in *Nature* (**427,** 29; 2004) last year showed that a flat stone should be tilted at 20° to the water's surface for optimal skipping. Now, Shin-ichiro Nagahiro and Yoshinori Hayakawa of Tohuku University in Japan back up this finding with theory. They model the impact of a stone on water using smoothed particle hydrodynamics, and calculate the angle where the velocity required for an onward bounce is minimized. Happily, for the peaceful coexistence of theorists and experimentalists at the waterside, they find it is always around 20°.

CHEMISTRY

Complex proteins

J. Am. Chem. Soc. doi: 10.1021/ja050304r (2005)

The functions of some proteins involved in cell signalling and the immune response are altered when small molecules, such as lipids and carbohydrates, are attached to an amino-acid side chain. Obtaining large amounts of these modified proteins for study has proved difficult because bacteria cannot make them.

A possible solution, offered by chemists from the University of Illinois in Urbana, involves an artificial amino acid containing aziridine, a three-membered ring. Thiols, the sulphur equivalent of alcohols, selectively react with this ring, forming a new bond. So far, the team has attached small thiol-containing molecules to peptides, but other methods may extend the technique's scope to longer proteins.

JOURNAL CLUB

Peter Dayan
University College London, UK

The director of UCL's Gatsby Computational Neuroscience Unit recommends some gainful reading.

Who of us lacking an anorak could admire an article with 60 equations and 15 instances of "piecewise linear"? But skip this paper — in the press at the *International Journal of Bifurcation and Chaos* — and you'll miss an important, if obscure, treat. It addresses a venerable question, which has long troubled me, in the neuroscience and psychology of decision-making: the role of neuromodulators.

Neuromodulators are an important class of transmitter. From a biophysical perspective, they seem to regulate general properties such as the excitability or 'gain' of their target neurons. Thus, many see them merely as widespread irrigants (or irritants, as our less charitable colleagues would have it).

But behavioural neuroscience studies suggest that neuro-modulators can play a key role at specific times in decision-making tasks. The transmitters regulate competition between populations of neurons that represent choices. This allows a subject to integrate noisy sensory information with past experience about rewards to make almost optimal decisions.

Previous attempts to reconcile the general function with the specific were a bit too heuristic to convince me. I buried my head in the sand or muttered jeremiads — complaining that the necessary tests were impossible.

However, this paper squares the biophysics with the behaviour. The researchers use a model to show that changing the gain of neurons is precisely the right thing to do to achieve optimal decisions. Along the way, they link classical and modern suggestions about the mechanisms of decision-making, an area under intense behavioural and electrophysiological scrutiny.

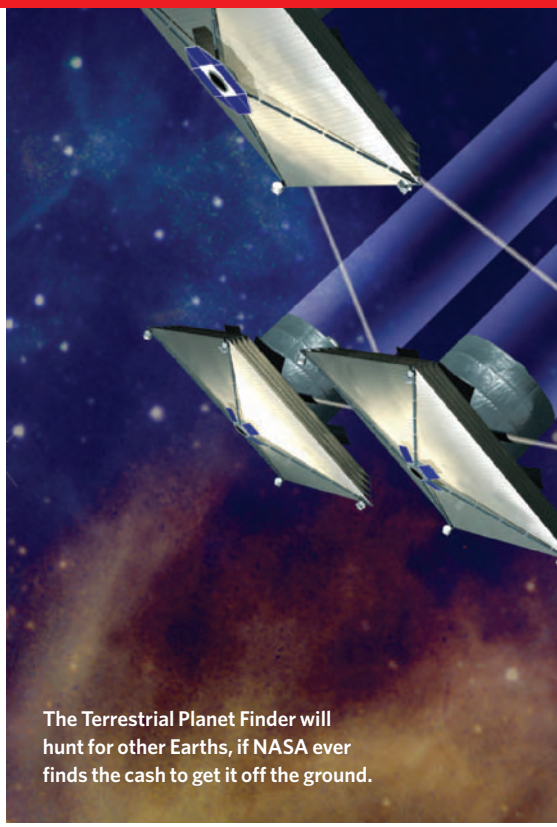
So buy your anorak, and let this theory reign.

www.math.nyu.edu/~ebrown/papers/simple_networks.pdf

©2005 Nature Publishing Group

NEWS

Planet hunters lose out to Hubble rescue



The Terrestrial Planet Finder will hunt for other Earths, if NASA ever finds the cash to get it off the ground.

Astronomers' elation at the prospect of a rescue for the Hubble Space Telescope is turning to dismay as the price of saving the venerable observatory becomes clear.

Last week, NASA turned in a revised budget plan to Congress that includes cuts and delays to several programmes, including the roving Mars Science Laboratory and searches for planets like Earth. The proposed cuts would also lead to belt-tightening in the Hubble project itself, where grants for guest observers would be reduced by an average of 13%.

Two planet-hunting projects — the Space Interferometry Mission (SIM) and the Terrestrial Planet Finder (TPF) — have been deferred until as yet unspecified dates. SIM would orbit the Sun, measuring stars' positions and trying to detect planets similar to Earth. The TPF, which consists of two space-based observatories, would follow up SIM's findings between 2014 and 2020 with detailed spectral analyses.

Both projects have been scaled back before. SIM is currently undergoing another redesign in an attempt to keep its cost below $1.2 billion. One possibility, says Charles Beichman of the Jet Propulsion Laboratory in Pasadena, California, who is on the SIM and TPF teams, is to advance its launch to 2010, because shorter development times often reduce costs. But that now looks unlikely.

The TPF was among the top recommendations of a National Research Council panel that set priorities in 2000 for the next decade of astronomy. But it is fraught with technical challenges — the panel called it "the most ambitious science mission ever attempted by NASA".

Deferring SIM and the TPF would be a blow to NASA's planetary search programme, which is already struggling with near-term projects. A plan to supplement the 10-metre ground-based optical Keck Telescopes, on Mauna Kea in Hawaii, with small 'outrigger' telescopes to detect Uranus-sized planets has been bogged down for years by battles with local cultural-rights groups (see *Nature* **417,** 5; 2002). And recent budget cuts have delayed the launch of the Kepler mission to look for very distant planets by eight months to June 2008.

Ground-based telescopes have found 155 planets around other stars and the European Space Agency lists the search for such 'extra-solar' planets among its priorities for the next decade. This leaves Beichman and others worried that NASA is neglecting a promising field.

Some are also anxious that the cost of a shuttle rescue mission to Hubble is squeezing — or giving NASA an excuse to squeeze — other projects. Most astronomers would support saving Hubble if money were no object. But an increasing number agree with Nobel laureate and astrophysicist Joseph Taylor of Princeton University, New Jersey, who told a congressional committee: "I do not favour such a plan if it would require major delays or reordering of NASA's present science priorities."

David Black, chairman of the American Astronomical Society's public-policy committee, says the balance of opinion has shifted towards the idea that Hubble isn't worth the sacrifice of future missions. Yet he says the telescope is only one of NASA's financial

Large genomic differences explain our little quirks

COLD SPRING HARBOR, NEW YORK

When the finished sequence of the human genome was unveiled last year, biologists said that it told a story of harmony for the human family. Every one of us, it turns out, shares 99% of our DNA with all the other people on Earth. But it's our differences that really fascinate us. And at last week's annual genome meeting in Cold Spring Harbor, New York, scientists revealed a wealth of data indicating a surprising conclusion about human diversity — much of it might be explained by large structural differences between individual genomes, not by tiny differences in individual genes.

PHOTOLIBRARY

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

People's genomes just don't line up.

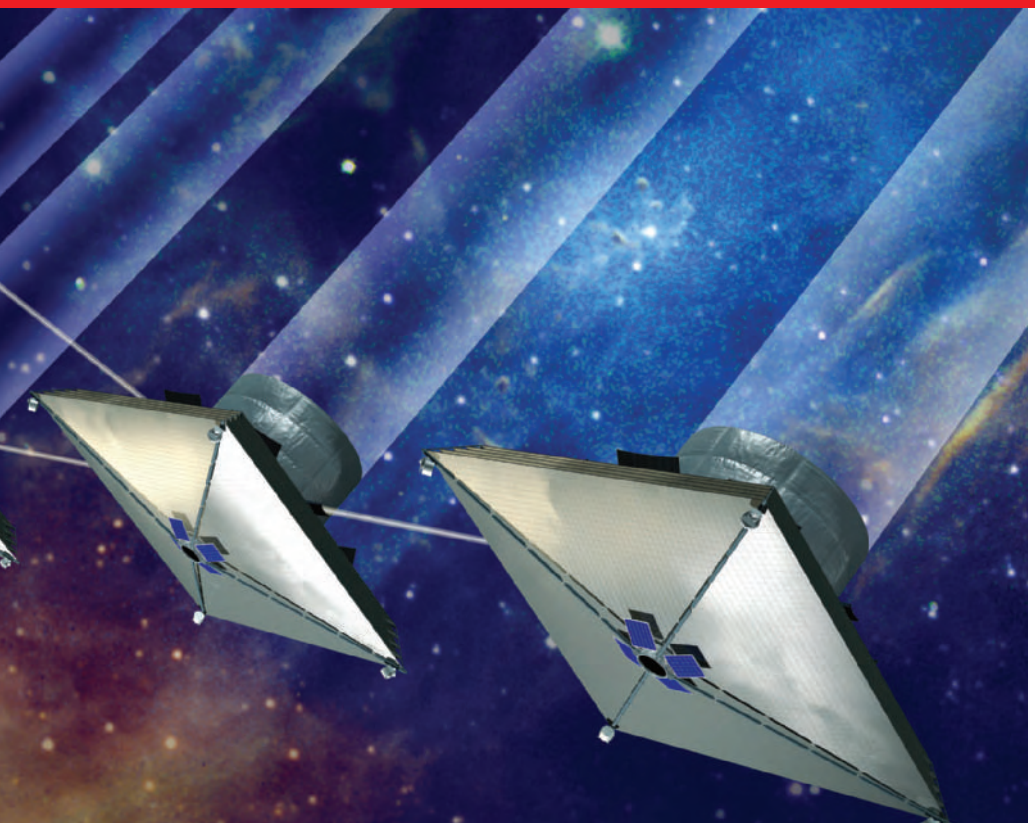
Biologists used to think that our genomes all had the same basic structure — the same number of genes, in roughly the same order, with a few minor differences here and there in the sequence of DNA bases. Now, technologies comparing whole human genomes show that this picture is incomplete.

Two years ago, a group of researchers led by Michael Wigler at Cold Spring Harbor Laboratory found the first evidence that some of us have more copies of certain genes than do others (R. Lucito *et al. Genome Res.* 13, 2291–2305; 2003). And at last week's meeting, Evan Eichler of the University of Washington in Seattle reported that this is just the beginning: not only do we carry different copy numbers of parts of our DNA, we also have varying numbers of deletions, insertions and other major rearrangements in our genomes.

In fact, Eichler found at least 297 places in the genome where different individuals have different forms of these major structural variations. At these spots, some of us might carry a major deletion, for example, or an extra hundred bases of DNA.

But do such differences mean anything?

©2005 Nature Publishing Group



NASA

headaches. Equally to blame are congressional 'earmarks' for pet projects, and cost overruns in many of its science programmes.

The latest example of that is the James Webb Space Telescope, due to launch in 2011. In early May, project managers learned of a $1-billion overrun that has raised its price to a whopping $3.5 billion. No obvious solution is in sight, says project scientist John Mather of the Goddard Space Flight Center, based just outside Washington DC. Shrinking the telescope is not acceptable to astronomers. "What comes next we don't know," says Mather.

The new NASA administrator Michael Griffin presented the revised budget to a Senate appropriations panel last week. "NASA cannot afford everything that is on its plate," he says.

Griffin's solution is to cut projects in the early stages of development. He also made it clear that: "In order to service the Hubble Space Telescope and provide for a safe deorbit, NASA will need to defer work on more advanced space telescopes."

He did have some good news — a NASA team thinks it can cut the number of missions required to complete the International Space Station from 28 to 18, and it is still trimming. At half a billion dollars per shuttle flight, such savings are very welcome. ■

Tony Reichhardt

Here, too, fresh evidence paints an intriguing picture. In January, scientists at the Iceland-based company deCODE Genetics found a long inversion — a stretch of DNA that is flipped around backwards — that is common in Europeans, but not in Asians and Africans (H. Stefánsson *et al. Nature Genet.* 37, 129–137; 2005). They also found that women who have this inversion bear more children than those who don't — a classic sign that the inversion confers an evolutionary advantage.

At the Cold Spring Harbor meeting, scientists presented more evidence that structural differences are important in human evolution. Duc-Quang Nguyen, a postdoctoral fellow in Chris Ponting's laboratory at the University of Oxford, UK, reported an analysis of areas where there are different numbers of copies of DNA stretches. Nguyen found that natural selection is actively working on these genes.

What's more, he found that many of these genes belong to groups that seem to help us interact with our environment. For instance, many work in the immune system, and affect how we fight off disease. These are exactly the sort of genes that could explain our diversity — why some of us get asthma when exposed to air pollution, or why some of us can eat plenty of cheeseburgers without gaining weight.

"We knew these variations existed, but this year we're asking, do they matter?" says Ewan Birney, head of bioinformatics for the European Molecular Biology Laboratory, based in Cambridge, UK. "The answer seems to be yes."

We're still one human family, of course; but our DNA landscapes are a lot more varied than we had thought. ■

Erika Check

ON THE RECORD

"We decided we could alter the discovery date for the opening of the movie."

Palaeontologist John Horner explains how he misled the press about his *Tyrannosaurus rex* discovery in order to promote *Jurassic Park III*.

"Antiretroviral drugs are expanding the AIDS epidemic."

South African maverick Matthias Rath takes out an outrageous full-page advert in *The New York Times* to accuse drug companies and the United Nations of genocide.

"'Because it's there' was reason enough to conquer Everest, but is it enough for scientific projects?"

Australian health minister Tony Abbott calls for heavier regulation of scientists.

SCORECARD

▲ **Orbiting tourists**
Dennis Tito's travel agents open a Tokyo office — so those with a yen for space should head for Japan.

▶ **Mars Express**
Don't expect the next 'Water on Mars' headline just yet. The first of the orbiter's three water-divining radar booms has unfurled, but snags delay the next two.

▼ **Fusion project**
Negotiators discussing where to build the ITER fusion reactor vehemently deny reports of an agreement. That July deadline is looming.

NUMBER CRUNCH

$4 million What USAID spends on bednets, drugs and insecticides to combat malaria.

$10.5 million What USAID spends on research into possible malaria vaccines.

$65.5 million What USAID spends on other costs, such as technical advice and consultants.

Estimates from Roger Bate, US director of Africa Fighting Malaria, in his Senate testimony (see page 257).

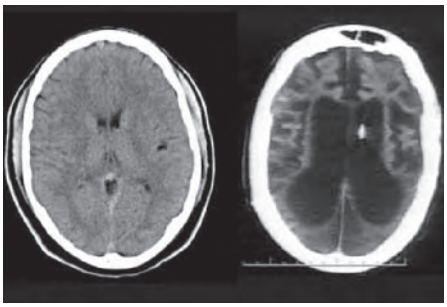
SIDELINES

©2005 Nature Publishing Group

Ethicists urge caution over emotive power of brain scans

WASHINGTON DC

Images have power — and Paul Root Wolpe knows just how persuasive they can be. A bioethicist at the University of Pennsylvania, Wolpe found himself confronted by angry members of the public during the debate over the fate of brain-damaged patient Terri Schiavo earlier this year.



P. R. WOLPE

A scan of Terri Schiavo's brain (right) shows just how damaged it was compared with a normal one.

People were outraged that Schiavo's husband wanted to remove her feeding tube, Wolpe recalls. But when he showed them computed tomography scans that compared Schiavo's atrophied brain with a normal brain, they changed their minds. "People paused, and they said: 'Maybe Terry isn't there'." Wolpe explains.

Last week, scientists and ethicists met in Washington DC to discuss the growing power of brain images to sway public opinion in areas such as medicine, crime and human rights. At the workshop, organized by a consortium of groups including the Dana Foundation, which funds research in neuroscience, they warned that neuroscientists must learn to use such emotive images responsibly, and to be more honest about the limitations of their work.

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Researchers are flocking to the field of brain imaging, and they are probing brain activity not only in diseases such as depression, but also in basic processes such as consciousness, decision-making and deception. Imaging

©2005 Nature Publishing Group

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Mind games: these composite scans show areas of the brain that are activated when someone doesn't tell the truth — but should such images form the basis of a lie detector?

KRT/NEWSCOM

studies are almost ten times as popular now as they were a decade ago (J. Illes *et al. Nature Neurosci.* **6,** 205; 2003).

The images have been a hit with the public too, because they are easier to understand than complex information about neurons, synapses and neural architecture. But this is causing concern about how the pictures might be used.

"These images are quite seductive," says Marcus Raichle, one of the pioneers of brain imaging at the Washington University School of Medicine in St Louis, Missouri. "It's intuitively easy to relate to a picture, and that's both good and bad."

Because brain images seem to tell a cut-and-dried story, the public often accepts research findings long before the investigators themselves feel they really understand what the results mean. Daniel Schacter at Harvard University, for instance, has used functional magnetic resonance imaging (fMRI) to image the brain regions that are activated when we remember things. He found that when we have 'false memories' — when we do our best to tell the truth but remember incorrectly — the images show different patterns of brain activation from those seen for true memories (S. D. Slotnick and D. L. Schacter *Nature Neurosci.* **7,** 664–672; 2004).

Schacter says his work is far from definitive, but that probably won't stop criminal investigators adopting the technique as a 'lie detector' test. "There's enormous public pressure to use these things," adds Wolpe.

Neuroimaging studies also raise issues relating to fundamental human freedoms. For instance, fMRI research supports the idea that we use emotions and reason when we make decisions. But people with certain kinds of brain damage rely more exclusively on reason — and often make what most would see as disastrous errors of judgment. So images of decision-making pathways could be used to label certain people as sociopaths.

But this could violate long-held societal standards, such as the 'freedom of thought' guaranteed in the First Amendment to the US Constitution, says Stacey Tovino, a lawyer at the University of Houston Law Center in Texas. "The privacy issues are very difficult," she notes.

The issues related to brain imaging are part of a larger debate about the implications of neuroscience research (see *Nature* **433,** 185; 2005). In 2002, the Dana Foundation sponsored a meeting that is widely regarded to have given birth to the field of neuroethics. Since then, the science has advanced rapidly — and the ethical problems have grown just as fast.

Everyone at last week's meeting seemed to agree that neuroscientists now have a responsibility to explain the limitations of brain-imaging studies to the public. For example, most non-scientists don't understand that the images in research publications are generally composites of data from many individuals, not from a single person. Or that the images are not photos of what's actually happening in the brain — they are highly processed representations of data.

Some delegates went further, arguing that scientists have an additional obligation to be more rigorous in their research. Many neuroimaging studies focus on a small group of people, and their findings aren't confirmed by other investigators.

"We need to have a new sense of accountability and reproducibility," says Judy Illes, director of the neuroethics programme at Stanford University. After all, she argues, with the public so eager for information about the brain, the least scientists can do is deliver the most accurate picture they can. ■

Erika Check

©2005 Nature Publishing Group

GETTY

PRIMATES TO PEOPLE
Viruses in Africa may be regularly making the leap between species
www.nature.com/news

IMAGES TAKEN BY RIT & JHU. COPYRIGHT RESIDES WITH THE OWNER OF THE ARCHIMEDES PALIMPSEST

SNAPSHOT

Eureka moment as X-rays slice through forgery

Synchrotron X-rays have been used to reveal long-lost writings from the tenth century.

The *Archimedes Palimpsest* is the earliest transcription of the theories of the legendary Greek mathematician, and the original text detailed his method of mechanical theorems. In the twelfth century the parchment was scraped clean and reused — a common practice to conserve the expensive material. It was turned into a prayer book, but the original writings survive as ghostly outlines beneath the prayers.

They can mostly be read using ultraviolet light, but some pages are also obscured by twentieth-century forgeries of medieval art (see picture). So researchers at the Stanford Synchrotron Radiation Laboratory in Menlo Park, California, used synchrotron X-rays on the piece. These cause the iron in the ink to fluoresce , creating a sharp image (see inset) of the priceless text.

US aid agency grilled over malaria funds

One of the world's largest aid agencies, the US Agency for International Development (USAID), has been hauled before a congressional committee. It faces allegations that its malaria programme is neither financially transparent nor effective. If the agency doesn't start being more open about its spending, it could even see some of its funding being given to other malaria groups.

Congress decided to investigate the agency after an article on malaria spending by Amir Attaran, a law professor at the University of Ottawa, Canada, appeared in *Nature* (**430,** 932–933; 2004). A Senate subcommittee heard the results at a hearing on 12 May.

USAID's malaria budget has quadrupled since 1998 to $90 million this year. The committee heard that several investigators trying to find out where that money goes had been unable to get firm figures from the agency. Senator Sam Brownback (Republican, Kansas) complained that his requests for detailed spending breakdowns had been met by "ad hoc responses with vague descriptions and math that doesn't add up". He also accused the agency of using "deceptive rhetoric and distorted science" to defend its actions.

Roger Bate, the US director of Africa Fighting Malaria, a non-profit advocacy group located in South Africa and the United States, presented his best estimates based on the "flawed data" he got from USAID. In 2004, he says, the agency spent just 5% of its funds on commodities such as bednets, drugs and insecticides, and 13% on vaccine research. The rest of the money went on other costs such as technical assistance, meetings and consultancy fees.

Michael Miller, deputy assistant administrator for global health at USAID, defended the agency's record. He argued that funding and planning have reached adequate levels only recently. "Any summary judgement of the progress or lack of progress is simply premature," he said.

Attaran testified that his research suggests USAID malaria funds went mostly on "American consultants, American experts, and very highly paid American 'non-profit' organizations". But Kent Hill, USAID's acting assistant administrator for global health, told *Nature* that "USAID prioritizes hiring local staff and consultants", and building local infrastructure.

Brownback told *Nature* that the US public "deserves nothing less than total transparency". He has introduced a Senate bill that would require USAID to describe its spending on a public website, and to use half of its budget for purchasing drugs and other commodities.

But Carlos Campbell, who heads MACEPA, an African malaria-control programme, cautioned that Brownback's bill may go too far. Technical assistance is often valuable to countries planning control programmes, he noted. "I ask you to not move everything to commodities, but to keep a balanced view," he said.

In wrapping up, Tom Coburn (Republican, Oklahoma), a physician and chairman of the Senate hearing, pledged to pressure USAID further. "I'm going to find out where every penny is spent," he said. If the agency doesn't cooperate, he warned, malaria funds could be stripped from it and transferred to the Global Fund to Fight AIDS, Tuberculosis and Malaria, an international group set up in 2002. ■

Declan Butler

©2005 Nature Publishing Group

SPECIAL REPORT

Taking on the cheats

The true extent of plagiarism is unknown, but rising cases of suspect submissions are forcing editors to take action. **Jim Giles** reports.

A. MACDONALD

The fight against plagiarism is about to take a decisive turn. Academic publishers have told *Nature* they hope that software designed to catch cheating students could soon be used to unmask academics who plagiarize other researchers' — or their own — work.

Big publishers such as Elsevier and Blackwell, which between them publish more than 2,500 journals, have been prompted to act by reports that plagiarism is becoming more common. "We're hearing about it more frequently from editors," says Bob Campbell, president of Blackwell Publishing in Oxford, UK.

Self-plagiarism, in which authors attempt to pass off already published material as new, is a particular problem. In an increasingly competitive environment where appointments, promotions and grant applications are strongly influenced by publication record, researchers are under intense pressure to publish, and a growing minority are seeking to bump up their CVs through dishonest means.

The extent of the problem is hard to assess. Defining plagiarism is not straightforward (see 'Where to draw the line?', below), and measuring the incidence of even the most clear-cut cases is difficult. Studies in certain fields have estimated that anything up to 20% of published papers contain some degree of self-plagiarism (see 'How common is plagiarism?', opposite). This may not be representative of basic research, but no rigorous, multidisciplinary study has ever been conducted.

And although most cases are never discovered, almost all of the editors and publishers contacted by *Nature* agreed that self-plagiarism is on the rise. "Editors are noticing many more cases," says Scott Dineen, director of editorial services at the Optical Society of America, which publishes ten journals. Last month, the increase prompted the society to issue an editorial statement on its commitment to expose plagiarism[1].

The advent of antiplagiarism software, such as that used by universities to check student essays, means that editors and publishers finally have a practical way to tackle the problem. Online services check essays against massive stores of documents generated from web trawls and purchases from media outlets. Supervisors can see which parts of the essays seem to be plagiarized and where the copied material comes from.

Adapting such technology for use with academic papers should be easy, say publishing experts, as the software could be bolted onto the online systems that publishers use to manage peer review. The system would work in the background and editors would only be aware of it when it flagged up a suspicious degree of overlap, which they could then check out in detail.

"We see it as an idea with great potential," says Campbell. "It will eventually be part of the editorial office."

ArXiv, the popular physics preprint server based at Cornell University in Ithaca, New

Where to draw the line?

Determining what constitutes plagiarism is tricky. Few scientific organizations have quantified the fraction of material that can be legitimately reused between papers, and few researchers have heard of those rules that do exist. So although plagiarism is relatively simple to define in a qualitative sense, it can be extremely difficult to rule on in practice.

Statements on plagiarism usually define the act as attempting to pass off someone else's work as your own. Duplicate publication, or self-plagiarism, occurs when an author reuses substantial parts of their own published work without providing the appropriate references. This can range from getting an identical paper published in multiple journals, to 'salami-slicing', where authors add small amounts of new data to a previous paper.

When large chunks of text have been cut-and-pasted, such definitions work well. But researchers routinely commit minor plagiarism without dishonest intent, such as reusing parts of an introduction from an earlier paper. To help editors resolve these cases, some journals set an upper limit for the amount of text that can reused, usually about 30%.

But what should an editor do when a paper looks similar to one already published, but does not contain chunks of text that have obviously been copied? This technique, dubbed 'intelligent plagiarism', is likely to evade detection tools that simply compare strings of text, although including sets of data in the comparison may help to flag up suspect cases.

Generally, the problem can only be resolved by editors studying the two papers, talking to the authors and making a personal decision on whether misconduct has occurred. "Plagiarism is always a human judgement," says Fintan Culwin, a plagiarism-software expert at London South Bank University.

Such cases can also be extremely time-consuming to investigate; one reason why data from the Committee on Publication Ethics, a UK-based group of biomedical journal editors, suggest that many allegations remain unresolved a year after they were first made. **J.G.**

©2005 Nature Publishing Group

CLIFF QUIVERS WARN OF COLLAPSE
Seaside cliffs may shake before they fall, giving hours of notice.
www.nature.com/news

York, is almost ready to deploy plagiarism detection software. Paul Ginsparg, the Cornell physicist who runs arXiv, acted after 22 plagiarized papers were discovered on the archive[2].

Physical exercise

Daria Sorokina, a computer science PhD student at Cornell, has tuned an established algorithm to look for any two documents that share at least six of the same words in a row. The system is already finding "plenty of awkward things", says Ginsparg. One test-run revealed a PhD thesis that shares large chunks of material with a paper posted to the archive three years earlier. So far, Ginsparg says, the tests have revealed a few thousand pairs of articles by different authors that have "excessive overlap".

Ginsparg plans to post all of the pairs on the arXiv website — without accusing the authors of wrongdoing — and ask the researchers involved to respond. He hopes that the results will help to refine the algorithm, which could then be used on new submissions to generate a warning if papers seem to overlap.

The world's largest scientific publisher — Amsterdam-based Elsevier, which publishes about a quarter of a million papers each year — has also decided to act. Last month, it initiated a year-long assessment of the various technology options.

Tools are now becoming available for individual editors and peer reviewers. One has been developed by Christian Collberg, a computer scientist at the University of Arizona, Tucson. He was asked to review a paper for a conference in 2003, and recalls carrying out a Google search to research it. "I found an earlier published version that had just been reformatted," he says. Reviewing for the same conference the following year, Collberg found another submission that had copied substantial parts of the author's own earlier work. "This really pissed me off," he says. "I spent time reviewing those papers."

Copycats

In response, Collberg began work on what has become the Self-Plagiarism Detection Tool (SPlaT). Whereas publisher-run plagiarism detection services are likely to take years to set up, Collberg's software is available, free-to-use, and targeted at editors and peer reviewers.

The software grabs papers from authors' websites and compares them to each other and to other manuscripts added manually, such as papers under review. Collberg declined to reveal details of what happened when he let SPlaT loose on the websites of 50 computer-science departments, but summary results released last month show that the software turned up more than one pair of conference publications with more than 50% common text and no reference to each other[3].

Student antiplagiarism services are another option. These offer the possibility of detecting plagiarism itself, not just duplicate publication. The firm behind the iThenticate software, for example, says its product is licensed by 5,000 institutions and that it can check documents against a database of more than seven billion pages for cases of suspicious overlap. Because this store has been generated in part from web trawls, it contains papers from authors' websites and some open-access journals. If a university subscribes to the service to tackle student fraud, researchers in those institutions can also use it when refereeing papers.

When *Nature* gave the software a test-drive, it did a good job of identifying areas where the same text had been legitimately used in different places, such as on authors' websites and in properly referenced quotes. But when a known plagiarism was submitted to iThenticate, it failed to turn up the original, even though it had been published in a high-impact journal.

The reason, says John Barrie, president of iParadigms, the California firm that developed the software, is that most online literature is locked behind subscription barriers and so cannot be added to the iThenticate database.

An overall solution to plagiarism, says Campbell, is likely to come when publishers collaborate on industry-wide detection systems. Preliminary discussions about this have already taken place at meetings of CrossRef. This company, based in Lynnfield, Massachusetts, collaborates with publishers to develop systems that allow researchers to search across journals from many different companies.

Such a system could, in theory, catch almost all instances of direct plagiarism, although it will take several years to set up, even if negotiations go smoothly. In the meantime, say editors, plagiarized papers will continue to creep into the literature as a minority of researchers dishonestly beef up their reference lists. After all, says one researcher who admitted to *Nature* that he occasionally neglects to mention papers that overlap with new publications, there is a perception that "the Dean can't read, but he can count". ■

How common is plagiarism?

A literature search on the word 'plagiarism' reveals numerous editorials bemoaning the problem, but little in the way of hard facts. The data are patchy, often anecdotal and rarely applicable to more than one field. To make matters worse, it's impossible to know how many cases evade detection.

One approach to gauging the frequency of plagiarism and other forms of misconduct is to search for notices of retraction. The results of one such trawl, which used biomedical literature in the PubMed database, were published in January and put the incidence of "recognizable fraudulent material" at less than 0.02% of all papers[4].

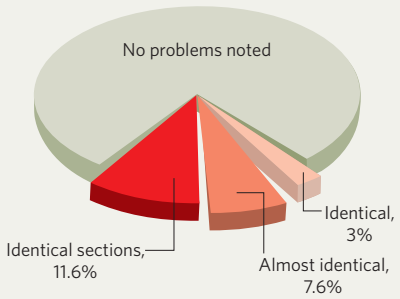
But surveys that compare individual papers report higher figures. Most of these focus on duplicate publication of clinical papers. Last year, for example, a trawl of 1,234 articles on anaesthesia and analgesia found that 5% were duplicates that did not reference the appropriate original[5].

A 2001 study of surgical journals put the figure even higher: nearly a quarter of articles published that year had some form of redundancy, and 11% were suspected to be dual publications (see chart). "Redundant publications must be recognized as a real threat to the quality and intellectual impact of surgical publishing," the authors were moved to conclude[6]. Duplicates have also been shown to cause meta-analyses to overestimate the efficacy of drugs[7].

The extent of the problem for most basic research is probably somewhere between these two extremes. Rigorous studies have not been performed on basic-research papers, so the problem may be going undetected. But commercial pressures may also encourage the duplication of papers that report positive findings about a new drug. In basic research, where the link between data and profits is less direct, the problem may be less common. **J.G.**

Cases of plagiarism

As found in 660 articles published in 3 top surgery journals



SOURCE: REF. 6

1. *Opt. Lett.* **30,** 813 (2005).
2. *Nature* **426,** 7 (2003).
3. Collberg, C. & Kobourov, S. *Commun. ACM* **48,** 88–94 (2005).
4. Claxton, L. D. *Mutat. Res.* **589,** 17–30 (2005).
5. von Elm, E., Poglia, G., Walder, B. & Tramèr, M. R. *J. Am. Med. Assoc.* **291,** 974–980 (2004).
6. Schein, M. & Paladugu, R. *Surgery* **129,** 655–661 (2001).
7. Tramèr, M. R., Reynolds, D. J. M., Moore, R. A. & McQuay, H. J. *Br. Med. J.* **315,** 635–640 (1997).

©2005 Nature Publishing Group

Bill Gates spurs disease research with $250 million

Microsoft chairman Bill Gates is to plough an extra quarter of a billion dollars into research on neglected diseases.

The money will be channelled through the Grand Challenges in Global Health initiative, which was set up to fund studies of 14 major scientific obstacles to the treatment and control of killer diseases such as malaria and tuberculosis.

The initiative, a joint project between the Bill & Melinda Gates Foundation, based in Seattle, Washington, and the US National Institutes of Health, was launched in 2003. With the new funds it will have almost half a billion dollars to distribute to research projects (see *Nature* 421, 461–462; 2003).

The first round of grants, chosen from 1,500 proposals received from more than 10,000 scientists in 75 countries, are expected to be confirmed this summer.

Gates made his announcement on 16 May when he addressed ministers from 192 states at the annual assembly meeting of the World Health Organization in Geneva.

www.grandchallengesgh.org

NASA/JPL

Pitting their wits: NASA researchers use a model to test ways of freeing their stranded Mars rover.

NASA team digs deep to ease that sinking feeling

Scientists at NASA have got out their buckets and spades in a bid to work out how to free Opportunity, the Mars rover that has been stuck in a sand dune since 26 April.

The researchers have placed a clone of Opportunity in an artificial dune at the Jet Propulsion Laboratory in Pasadena, California. When the dune was made of beach sand, a test rover easily got itself out, so the team switched to a mix that included clay powder, which more closely resembles the powdery low-traction Mars sand.

Their tests suggest that after spinning its wheels a bit, the rover should be able to get moving without sinking further into the sand. The plan could be put into action next week.

Poor countries slip further behind in health research

Low-income countries are publishing a smaller fraction of health-related scientific papers than they were a decade ago, say researchers at the World Health Organization.

The team, which looked at almost 3.5 million articles published in more than 4,000 journals between 1992 and 2001, found that output from the world's 63 poorest countries had dropped by about a tenth to just 0.3% of all health publications. Whereas middle-income nations such as China and Turkey have boosted their output by some 20–30%, just one poor nation — India — produced a substantial number of papers (G. Paraje, R. Sadana and G. Karam *Science* 308, 959–960; 2005).

Centenary stimulates top scientists' wish-list

What's the one thing you wish more people grasped about science? That was the question posed to some 270 leading scientists, whose responses have now been published for all to ponder on the Internet.

©2005 Nature Publishing Group

The results from the survey, commissioned by the online magazine *Spiked* to mark the centenary of the equation $e = mc^2$, reveal the desire of many scientists to show what sets the scientific method apart from other spheres of intellectual endeavour.

"I should teach the world that science is the art of doubt, not of certainty," reads the contribution from Frances Ashcroft, a physiologist at the University of Oxford, UK. "Science is the antithesis of faith, and of the popular view that science provides immutable theories and fixed facts about the world in which we live."

The survey was launched on 10 May with a panel discussion at the Royal Institution of Great Britain in London.

www.spiked-online.com/einstein

Rival bidders shape up in fight to run Los Alamos

The Los Alamos National Laboratory in New Mexico will almost certainly be run by an academic–industrial partnership when the current contract to manage the labs expires on 30 September.

The University of California, which has operated the giant nuclear-weapons research facility for more than 60 years, announced on 11 May that it is joining forces with an industrial team led by Bechtel to bid for the new contract.

Bechtel, a global construction and management firm based in San Francisco, will collaborate on the bid with two nuclear-industry firms: BWX Technologies of Lynchburg, Virginia, and the Washington Group International of Boise, Idaho.

The only other bidder — technology firm Lockheed Martin of Bethesda, Maryland — said on 12 May that it would work with the University of Texas on its proposal. Texas regents have set aside $1.2 million for the bid.

Tsunami silt brings dark prospects for coral reef

Last December's tsunami in southeast Asia blanketed the coral reefs of the Andaman and Nicobar Islands with thick layers of silt and sand. Removing it will require "huge effort and several years", say Indian oceanographers.

By shutting off sunlight, the sand will eventually choke the marine species that coexist with the corals, upsetting the intricate ecological balance, according to researchers who returned from a ship-based survey on 7 May. The reef, which stretches for 800 kilometres from Indonesia to Myanmar, is home to 200 coral species — second only to Australia's Great Barrier Reef in species diversity.

In a parallel survey using Global Positioning System satellites, Indian geologists found that the 9.3-magnitude quake that triggered the tsunami also shifted the Nicobar Islands southwest by about five metres and sank some of the Andaman Islands by as much as one metre. Officials say the satellite and coral-reef observations will continue on a long-term basis.

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Tidal damage: debris from last year's tsunami litters a coral reef off the coast of Thailand.

A. PATRICK/GAMMA

©2005 Nature Publishing Group

The brains trust of Tehran

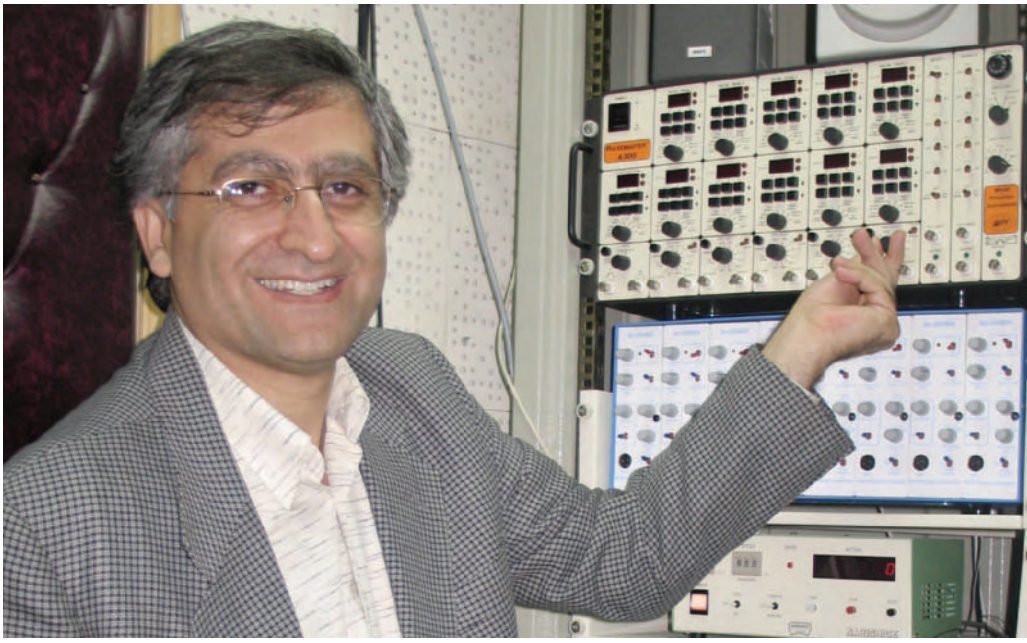
An Islamic theocracy ravaged by economic sanctions isn't an obvious place to seek a vibrant cognitive-neuroscience research group. Yet that's what **Alison Abbott** found on a recent trip to Iran.

In December 1996, 20 students gathered in a small candle-lit room in the north of Tehran. There they vowed to devote themselves to the study of the brain and mind, sealing their pledge in a document signed with their own blood. In a country with no textbooks or journals in cognitive neuroscience —and no teachers in the subject — this solemn ceremony symbolized extraordinary commitment.

At around the same time, Hossein Esteky abandoned a brief struggle to establish himself at the Tehran Medical School, where he had returned after completing a PhD in neurophysiology at the University of North Texas. He left for a two-year postdoc position at the RIKEN Brain Science Institute in Saitama, Japan, unsure if he'd ever be able to pursue his professional interests in his homeland.

Today, against the odds, Esteky heads a cognitive-neuroscience research group in Tehran that is beginning to make its mark internationally. Many of the team are former members of the 1996 study group, known as Niloufar — the Farsi word for lotus flower, and the street where its inaugural ceremony took place. Western scientists who have paid a visit to Esteky's group are effusive in their praise. "The laboratory is a wonder," says Patrick Cavanagh, a vision researcher at Harvard University.

Niloufar formed around a nucleus of three medical students, all coincidentally known as Arash, and each the product of the élite high schools run by NODET, the National Organization for Development of Exceptional Talents. NODET's extracurricular workshops, remembers one of the trio, Arash Yazdanbakhsh, "instilled in us the spirit of experimentation". The three have since vindicated NODET's backing — and proved themselves worthy bearers of the name of an ancient Iranian hero.



A. ABBOTT

"When I got to the Tehran institute, I found these eager students hanging around, and naturally we hooked up." — Hossein Esteky

Iranian mythology tells of Arash the Archer, who accepted a challenge from the attacking armies of Turan in the north. Shooting from Mount Damavand, Iran's highest peak, the country's border would henceforth be drawn where Arash's arrow fell. The Turans hoped to confine the Iranians to a narrow strip of land. But when Arash shot his bow, he turned into pure energy, propelling the arrow 2,250 kilometres to the Oxus River in Central Asia.

His three modern namesakes, working in a country devastated by the war with Iraq in the 1980s, and kept crippled by US-led sanctions, set to their neuroscience challenge with similar energy. Their first task was to expand the group, then locate suitable books to begin their studies.

Bookworms

Arash Fazl, a technical whiz, borrowed a computer in the Institute for Studies in Theoretical Physics and Mathematics (IPM), Iran's premier research centre, and set up a DOS-based Internet connection. The group used it to identify three books on cognitive neuroscience. NODET head Javad Ejei bought the texts for his protégés during a trip abroad. The arrow was in the bow.

As soon as their medical classes finished, Niloufar's members began their 'real' study, recalls Seyed Reza Afraz, who has been known as Arash since childhood, and today works in Esteky's group. "We divided the books among ourselves for efficiency," he says. "And each of us taught the others what we learnt from them."

To begin with, the going was tough. Afraz and his peers realized that they should have started with more fundamental textbooks. But within a year they were getting up to speed, and

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

H. SARBAKHSHIAN/AP

Under the gun: US sanctions are a way of life in Iran, but they hamper scientific research.

©2005 Nature Publishing Group

started to hanker after primary research papers. Scouring journal contents on the Internet, they wrote e-mails to researchers whose work sounded intriguing. Cavanagh and Irving Biederman, a cognitive neuroscientist at the University of Southern California in Los Angeles, were among the first to send their papers.

"I was really moved by the requests that came from young people wanting to get out of the dark," says Biederman. He and Cavanagh each began an intense scientific correspondence with the Niloufar group. The students devoured every word and began to do basic experiments in psychophysics. They subjected human volunteers to simple visual stimuli displayed on a computer monitor, and drew inferences about cognitive processes from their responses.

The group's first results were published as an abstract submitted to a 1998 neuropsychology meeting in Montreal, Canada — which the Iranians could not afford to attend. By then, a spin-off study group had formed in Esfahan, an ancient city 400 km south of Tehran.

Meeting of minds

But it was back at the IPM where things really began to take off. One day in 1999, some Niloufar students were invited to work on a project with Abdolhossein Abbassian, an IPM mathematician who wanted to model the function of the thalamus, a key relay station for information in the brain. "He looked at us one evening," recalls Afraz, "and said: 'I am a mathematician. To me, the thalamus is a black box with input and output. I want to see the real thing, and touch it'." So Afraz and his friends dissected a lamb thalamus for him. They were rewarded with a take-away pizza, a luxury in Tehran.

Having got a foothold in the IPM, the students began to haunt the place, borrowing computers and other facilities. It was a good time to get involved. Set up in 1989 to pursue Iranian strengths in theoretical physics, the IPM was trying to expand into experimental work. Esteky had returned from Japan in 1999 to take up a position at the Shaheed Beheshti University of Medical Sciences in Tehran. The IPM invited him to do some projects at the institute, where he and the Niloufar students immediately hit it off. "When I got there, I found these eager students hanging around, and naturally we hooked up," Esteky says.

Esteky invited those interested in his own field of visual perception to join his projects, and provided working space for others whose interests left them without a mentor in Iran. The austere corridors of the IPM were enlivened with music and chatter until after midnight, as the exuberant Niloufar gang combined partying and work.

Even with such enthusiastic staff, it was hard for Esteky to establish his research. He wanted to continue the work he had been doing in Japan, trying to understand how monkeys categorize visual objects and recognize faces.

Happily, Keiji Tanaka, vice-president of the RIKEN Brain Science Institute, donated electrophysiological equipment. But getting monkeys was a challenge. Esteky toured around the zoos in Iran, and eventually built up a colony of 12 macaques. Finding no commercial source of feed, he had to contract a local baker to make the required biscuits.

Fishy tale

A much-told story is of Esteky's first operation on the brain of a monkey, done with some students at one in the morning. The power failed after the skull was opened, and panic erupted. The building was new and the emergency lights didn't work. One of the students rushed for a tin of tuna. He inserted a makeshift wick into its oil and the monkey was looked after by tuna-oil lamp until power was restored.

Today, Esteky heads the IPM School of Cognitive Neurosciences, and has a dozen PhD students, many from the Niloufar and Esfahan study groups. IPM funds have purchased additional infrastructure, including a visual neuroscience lab kitted out for about US$500,000. But the problems of working in a country gripped by US sanctions continue. It took four years for Esteky to get one piece of equipment for analysing neuronal recordings, made by a single American company. Some journals have made it difficult for Iranian scientists to publish. And visas for travel to the United States are hard to come by.

Visitors are few — even though Esteky invites many foreign scientists. Those who have been bold enough to fly to Tehran have been impressed. Cavanagh was the first to visit, in 2002. Sightseeing plans were soon abandoned, as he was mobbed by students who kept him occupied with high-level discussions until he left. "They had enthusiasm, creativity and a scientific fervour that matched or even surpassed any other group I have visited," says Cavanagh. "I left exhausted and inspired."

Nancy Kanwisher, a specialist in visual perception at the Massachusetts Institute of Technology, visited last year. "I nearly wept when I understood what they were doing," she says. "It is cutting-edge science, addressing extremely hot questions, against formidable odds." She was particularly impressed by the female students, who were as ambitious as their male counterparts. Iran's religious leaders may be conservative in their attitudes to women, but such views don't prevail in the lab.

Esteky's biggest concern is that his students' talent and ambition will drive many of them abroad. Each of the original Arash trio has already shot his arrow well beyond the borders of Iran. Yazdanbakhsh and Fazl are completing their PhDs at Boston University; Afraz will start his PhD with Cavanagh in September.

Esteky won't pressurize anyone to stay, and says his job is to create the best possible working conditions so that returning is a viable proposition. As a student during the war with Iraq, he volunteered for medical duty and treated civilians who were victims of Saddam Hussein's chemical weapons. The experience made him determined to strive for peace, stability and scientific progress in Iran. Between the ninth and sixteenth centuries, he notes, Iran was a world leader in science: who says it can't one day be so again? ■

Alison Abbott is *Nature*'s senior European correspondent.

"We divided neuroscience textbooks among ourselves for efficiency, and each of us taught the others what we learnt from them." — Seyed Reza Afraz

A. ABBOTT

©2005 Nature Publishing Group

NATURE|Vol 435|19 May 2005

ESA/DLR/FUBERLIN/G. NEUKUM

Going underground

A new generation of planetary radar aims to look deeper than ever into some of the Solar System's most enduring mysteries. **Tony Reichhardt** gets ready for a trip to the interior.

Photographs of the martian surface are undeniably beautiful, but uncovering the truth behind the sinuous river beds and suggestive gullies is not easy. Spacecraft and rover missions have turned up plenty of evidence that water once flowed on Mars, but today the planet's surface is bone dry. There is water ice at the poles and hints of a frozen sea near the equator — but any large reservoirs of water are likely to be kilometres down where conditions are warmer. So the hunt for liquid water, which might provide a habitat for martian life, has gone underground.

In the coming months, an instrument called MARSIS (Mars Advanced Radar for Subsurface and Ionospheric Sounding) on Europe's orbiting Mars Express spacecraft could supply some long-awaited answers. If it works as planned, it will be a kind of divining rod for Mars, and could pave the way for future radar projects.

As with divining rods on Earth, the answers are likely to be ambiguous, and MARSIS may draw a blank. But for the moment, project scientists will be happy if it works at all.

In April 2004, the MARSIS team had to delay starting up their radar, the last of seven instruments on Mars Express, after concerns that the antenna's two 20-metre-long fibreglass booms might whip around and damage the European Space Agency's (ESA) craft after their release. A team including experts from the Italian Space Agency and NASA's Jet Propulsion Laboratory, who created the radar, ran computer simulations and ground tests to assess the risk. ESA went ahead with the deployment in early May, but a glitch as the first boom unfolded has resulted in more analysis and delay. Meanwhile, precious radar viewing time is being lost. As Roberto Seu, a MARSIS investigator at the University of Rome 'La Sapienza', admits: "It's been very, very frustrating."

Prime time

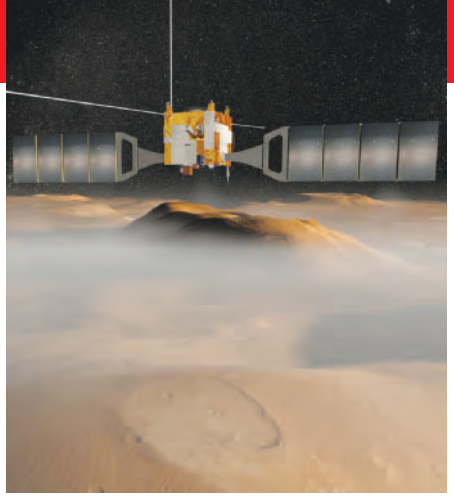
As the spacecraft's orbit shifts, the lighting and ionospheric conditions on Mars change, and the long wait has resulted in MARSIS missing much of its prime viewing window for this year. "We're very eager to get started," says Jeff Plaut, one of MARSIS's two principal investigators, who is based at the Jet Propulsion Lab in Pasadena, California. The radar scientists hope that Mars Express will be extended beyond its scheduled two-year mission, which would let the experiment continue early next year.

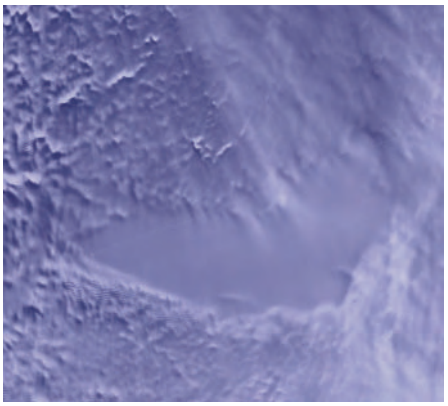
Meanwhile, a project called SHARAD (Shallow Subsurface Radar) is being readied for launch on NASA's Mars Reconnaissance Orbiter in August. If these experiments prove successful, radars could be attached to rovers destined for the martian surface, and sent to study Jupiter's ice-covered moon Europa or the interior of an asteroid.

Sounding radars such as MARSIS and SHARAD work by transmitting a radar pulse from orbit down to a planet, then analysing the signal that bounces back. Depending on the pulse frequency and the properties of the materials it strikes, the signal will penetrate a certain depth below the surface. In general, the lower its frequency, the deeper it will penetrate. Water ice is very transparent to radar. Dry soil can be too, depending on factors such as its electrical conductivity and porosity. Liquid water is a strong radar reflector. Tracking how much radar energy bounces back, how it's scattered, and how fast it moves through different patches of underground material can, with some coaxing, yield a rough map of subsurface geology.

If all goes well, MARSIS will begin scanning the planet's northern hemisphere this year. The radar's low frequency (1.3–5.5 MHz) allows it to penetrate, in principle, down to 5 kilometres. But such low frequencies result in low resolution — between 5 and 10 kilometre horizontally, and about 100 metres vertically. Any reservoir of groundwater smaller than that will not show up.

©2005 Nature Publishing Group





JPL/NASA/ESA

CANADIAN SPACE AGENCY RADARSAT-1 (1997)/NASA-GFSC

On the radar: Mars Express (top) has radar antennae that will search beneath the martian surface (left), hoping to discover subterranean reservoirs like Lake Vostok in Antarctica (above).

SHARAD's higher bandwidth and operating frequencies (15–25 MHz) mean that it won't penetrate as deep as MARSIS (about 1 kilometre), but it will have higher resolution — as little as 300 metres horizontally and 15 metres vertically. Data from the Mars Odyssey project suggest there is ice in the planet's shallow subsurface, which may be detected by SHARAD.

Broad spectrum

On Earth, geologists, engineers and archaeologists use portable ground-penetrating radar (GPR) units to hunt for everything from mineral deposits to buried artefacts. Mounted on aeroplanes, radars are also handy for exploring vast icy regions. They were used in 1996 to confirm the existence of Lake Vostok, a 10,000-square-kilometre reservoir hidden under 4 kilometres of ice in Antarctica.

Elsewhere, high-frequency orbiting radars have been used to map the surfaces of cloud-covered worlds such as Venus and Saturn's moon Titan, which are hidden from ordinary telescopes. But these radars barely scratch the surface. The only body apart from Earth to be probed by a deep sounding radar is the Moon — the Apollo 17 spacecraft carried an antenna that penetrated more than a kilometre below the dry lunar dust in 1972.

Roger Phillips, a geophysicist at Washington University in St Louis, Missouri, who worked on the Apollo experiment, recalls that interpreting the results was a nightmare. "We went into the Apollo thing pretty naively," he says. The radar data were full of echoes from surface features as well as underground rock layers, and real subsurface features got lost in the noise. It wasn't until they used stereo pictures of the surface taken from orbit that Phillips' team could factor out the radar echoes. And it was years before they were confident enough to publish their findings, which tentatively identified structures a kilometre down.

Eliminating confusing radar echoes will still be a problem on Mars, says Phillips, now co-leader of the team that is building SHARAD. But the Mars investigators know much more about the topography of the planet than the Apollo investigators did about the Moon in 1972. Using elevation data from NASA's Mars Orbiter Laser Altimeter and stereo pictures from the high-resolution camera on Mars Express — not to mention 33 years of improvements in computing power — they will be able to model the radar noise from surface features ahead of time.

But we still do not know how well MARSIS and SHARAD will work in practice. On Earth, a geologist would survey the subsurface rocks before using a sounding radar by mapping the local outcrops and mineralogy. But the Mars radars will be peering into terra that is mostly incognita. That makes their chances of penetrating the martian crust unpredictable. Some minerals with high metal content are good radar reflectors, and could prevent the radars from probing deep enough to find water.

Current models suggest that liquid water could exist 2–6 kilometres below the martian surface, but the radar signature for water is not unique, and other conductive materials, including certain clays and metallic minerals, could be mistaken for groundwater.

Phillips believes that the search for water will be difficult. But that is not MARSIS's sole purpose. Mapping the base of the planet's thick northern ice cap would be valuable for planetary geologists, for example — as would any clear picture of the subsurface.

Interior designs

John Grant, a planetary geologist at the Smithsonian Center for Earth and Planetary Studies in Washington DC, thinks that MARSIS will work over some areas but not others. He hopes to land a small, mobile GPR on the planet in the future and proposed one for NASA's next rover mission, the Mars Science Laboratory. It wasn't picked, but a successful MARSIS experiment could boost his chances when the next opportunity arises.

Grant's proposed GPR system would take very high-resolution (tens of centimetres) radar soundings to depths of 10–20 metres. When mounted on a rover, the device would take quick radar pulses of the subsurface at every place it stopped. Eventually it would build up a map of the geology underfoot. If tapping into water ice near the martian surface requires drilling, as most people think it will, radars could help future explorers pick the best site.

Europa is also in the sights of the radar gang. It is an intriguing target for planetary scientists because of the ocean thought to underlie its icy crust. However, any radar built to penetrate the thick ice would have to operate in Jupiter's strong radiation environment.

Even more exotic is a proposal from a team led by Erik Asphaug of the University of California, Santa Cruz, designed to probe the insides of an asteroid. The concept lost out in NASA's last competition for mid-sized planetary missions, but is likely to be submitted again. Earth-based radars have already been used to scan the outside of asteroids. Deep Interior, as it is called, would use a sounding radar in close orbit around an asteroid to take a three-dimensional image similar to a medical ultrasound. Combined with active experiments such as NASA's Deep Impact mission, which will blast an artificial crater in a comet this summer, it could provide the most complete picture of an asteroid yet.

But for all its promise, sounding radar still suffers from a lack of respect, or perhaps awareness, among planetary scientists. It's a complicated technology, and interpreting the data is tricky. "Most planetary scientists don't believe in radar," admits Seu. But MARSIS investigators are hoping that their experiment will soon change that. And all they really need is one Lake Vostok. ■

Tony Reichhardt writes for *Nature* from Washington DC.

©2005 Nature Publishing Group

NATURE|Vol 435|19 May 2005

It's the ecology, stupid!

Stem cells are engaged in constant crosstalk with their environment, biologists are fast realizing. So the emerging field of regenerative medicine is now wrestling with the ecological concept of the niche. **Kendall Powell** reports.

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Linheng Li is learning to think like an ecologist. His study subjects put down roots near sources of nourishment and depend on other living things in their environment to thrive. But Li doesn't have mud on his boots. The 'species' he studies are stem cells and the 'ecosystem' is bone marrow.

Within the anatomical forest of the marrow, Li's stem cells occupy specific niches — a term borrowed from ecology. An organism's ecological niche is a definition of where it lives, what it does, and how it interacts with its environment. Alter that environment, and the consequences for the organism can be dire. Conversely, if you take an organism and deposit it in an alien ecosystem, all hell can break loose.

At the Stowers Institute for Medical Research in Kansas City, Missouri, Li thinks in a similar way about stem cells. For example, in aplastic anaemia, stem cells are unable to produce sufficient blood cells, even though they look normal. Something about the cells' microenvironment in the bone marrow may be awry, argues Li. "It's like the soil being damaged," he says. Similarly, just as an introduced species can run amok in its new environment, stem cells placed in the wrong tissue in the body might conceivably form a malignant tumour.

Clearly, you can't hope to understand a woodland flower's niche in the forest by examining a specimen grown in a pot. And biologists are realizing that they are missing an important part of the picture by studying stem cells in Petri dishes. "Thinking of stem cells in isolation can be productive," says David Scadden, co-director of the Harvard Stem Cell Institute in Boston. "But it falsely simplifies what is a single component of a much larger, more complex system."

Indeed, this jungle of stem cells, support cells, protein scaffolds, blood vessels and biochemicals may be every bit as complex as a forest ecosystem. And teasing apart the 'ecological' interactions involved may be crucial if stem cells are to fulfil their therapeutic promise, and eventually be deployed to grow specialized cells and tissues to replace those lost to injury or disease.

So far, scientists have begun to define stem-cell niches for just a handful of tissues. Li's and Scadden's groups, for example, have partially defined the niche of blood-forming haematopoietic stem cells in bone marrow. Biologists suspect that the signalling pathways that help to define different stem-cell niches share common characteristics. But the details are likely to vary from tissue to tissue.

No place like bone

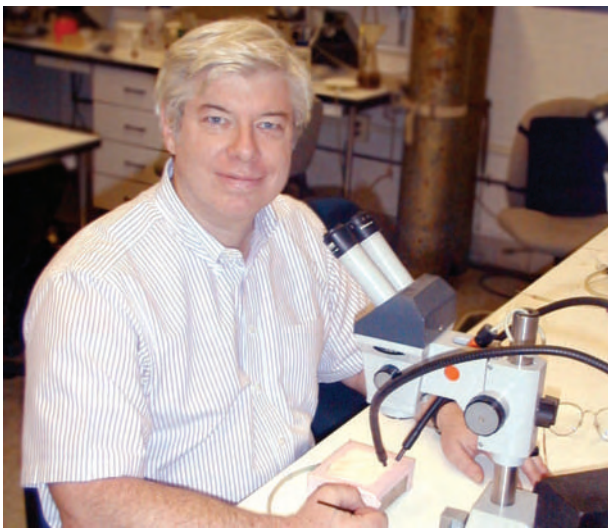
Researchers working with haematopoietic stem cells first proposed that the cells responded to their microenvironment almost four decades ago[1]. "They realized there was something special in the bone marrow — that it was an idyllic hideaway where stem cells could live for ever. As they leave the region, they are doomed to differentiate," says Haifan Lin, a stem-cell researcher at Duke University Medical Center in Durham, North Carolina.

But at that time, biologists lacked the molecular tools to reveal why bone marrow is a safe haven for its stem cells. They needed cell-surface markers that would distinguish stem cells from the other cell types present. And without gene knockout technologies, they were unable to eliminate components of the bone marrow environment to find out which ones were crucial for stem-cell survival.

When such tools appeared in the 1990s, they were first applied to a simpler system, in which it is easier to track the fates of individual cells. The gonads of fruitflies contain germline stem cells that give rise to sperm or eggs. These stem cells are held in sharply defined microenvironments, surrounded by just a few cellular neighbours.

Around the turn of the millennium, two research groups showed that a tight pocket of support cells physically holds on to the stem cells, lining them up to divide properly. The oriented divisions ensure that one daughter cell remains in this pocket, where it receives specific biochemical signals that maintain its 'stemness'[2,3]. The other is pushed out to start moving down the developmental pathway that gives rise to sperm or eggs.

"The stem cells may just be following along, not really calling the shots," suggests Allan Spradling, director of the Carnegie

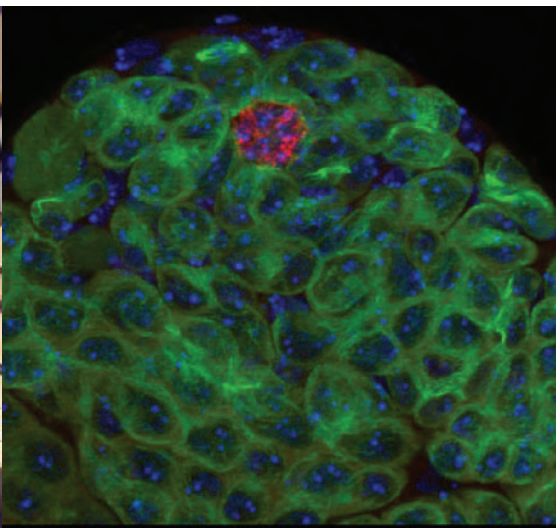


K. BELSCHNER

Allan Spradling was one of the first to draw parallels between stem-cell biology and ecological niches.

©2005 Nature Publishing Group

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS



Germline stem cells in fruitfly testis surround a 'hub' of cells (red) that direct stem-cell renewal.

L. JONES

Institution's Department of Embryology in Baltimore, Maryland. If the germline stem cells are removed, other cells — including those that have begun to differentiate — move in to take their place[4]. Similar dynamics can operate in ecosystems, if one species is wiped out.

In their pioneering papers[2,3], Spradling's team and the second group, led by Margaret Fuller of Stanford University in California, equated the term 'niche' with the microenvironment in which stem cells sit. Others have followed this convention. But from a strict ecological perspective, this is a misnomer — it's more of a nook, than a niche. In ecology, a niche describes a living thing's specific interaction with its environment, more like a job description than an address.

Regardless of the way in which the terminology is used, it's becoming clear that 'ecological' interactions are central to stem-cell biology. "The fly research has laid the groundwork," says Leanne Jones, formerly a postdoc with Fuller, now running her own lab at the Salk Institute for Biological Studies in La Jolla, California. The next challenge is to tease apart the ways in which the various types of stem cells found in adult mammalian tissues interact with their microenvironments.

The best-studied mammalian adult stem cells are in skin, intestine, bone marrow and a few specialized regions of the brain where new neurons form. In each case, specific support cells seem to hold stem cells in place, and help to direct the production of daughter cells that differentiate to form specialized cells and tissues.

In the bone marrow, for instance, Li and Scadden have shown that cells called osteoblasts lining the inner bone surface are closely associated with haematopoietic stem cells, encouraging their division and regulating their subsequent differentiation into the precursors of blood cells[5,6]. In a brain region called the hippocampus, neural stem cells take their cues both from cells called astrocytes and from the endothelial cells that line blood vessels[7,8]. Generally speaking, it seems that endothelial cells encourage stem cells to stay put and renew themselves[8], whereas astrocytes say 'become a neuron'[7].

Missing a trick: studying stem cells in culture overlooks interactions with their environment.

S. HUFFAKER/GETTY IMAGES

Researchers are now trying to tease apart the biochemical signalling going on between adult stem cells and their support cells. It seems that these often involve the same molecular pathways that direct tissue generation during embryonic development — such as the Wnt, Notch and TGF-β/BMP pathways familiar to cellular developmental biologists. "Maybe a stem cell is like the embryonic stage and is still able to respond to at least one of these pathways," says Li.

In some cases, stem cells are physically anchored to their support cells; here, biochemical signalling may require physical contact between cell-surface receptors. For instance, osteoblasts bear a membrane protein called jagged 1, which fits into and activates a receptor called Notch carried on the surface of haematopoietic stem cells. This stimulates the Notch pathway, which encourages the stem cells to keep dividing to renew themselves[6].

Location, location, location

In other tissues, such as the skin, support cells have a looser association with their stem cells. And in many cases, cellular communication is by diffusible biochemicals. The blood supply is also an important part of the microenvironment — many stem cells have blood vessels nearby, which may alert them to the physiological state of the whole organism. Some stem cells, such as those in bone marrow, can even migrate into blood vessels, circulate around the body, and return home when needed.

Thinking like an ecologist can help when trying to understand such complex behaviour. Neural stem cells, for instance, stay put in the hippocampus, but their relationship with blood vessels may be every bit as important as the relationship of plants with streams and rivers. "Sitting right up next to a blood vessel versus just being nearby may make a huge difference to a stem cell," says Fred Gage, whose lab at the Salk Institute has led the way in characterizing neural stem-cell niches.

Gage's group has also found that mammalian neural stem cells are not such passive players in their microenvironment as the fruitfly germline stem cells studied by Spradling and Fuller. Andrew Wurmser, a postdoc in Gage's lab, has found that neural stem cells can, under certain conditions, give

©2005 Nature Publishing Group

rise to endothelial cells[9]. This hints that, if the need arises, they may be able to populate their microenvironment with the support cells that they need to thrive.

Researchers led by Elaine Fuchs of Rockefeller University in New York have added further weight to the idea that some stem cells can create their own microenvironments. She has isolated individual epithelial stem cells, which give rise to skin and hair, and cultured them in the lab. When she subsequently grafted cells from these cultures into the skin of mutant hairless mice, they gave rise to hair follicles containing stem cells and support cells[10]. The nude mice grew tufts of hair.

Many researchers want to work out whether adult stem cells can be induced to widen their horizons from their usual niche, and give rise to different cells and tissues. The idea that some stem cells can 'transdifferentiate' in this way has been hugely controversial, with many early findings now dismissed as artefacts. "If you put a haematopoietic stem cell into a neural stem-cell niche, does it become a neural stem cell, or will it respond differently? We still don't know," says Wurmser.

Transdifferentiation remains an important question. For some diseases that might, in theory, be treated using stem cells, no appropriate natural stem cell has yet been identified — type 1 diabetes, where the pancreas loses its insulin-secreting β-cells, is the best-known example. Many researchers wonder whether an improved understanding of stem-cell microenvironments will provide clues about how to send stem cells down different developmental pathways, or reactivate quiescent stem cells. Perhaps a haematopoietic stem cell will give rise to a β-cell if placed in a pancreas-like environment. Maybe one day it will even be possible to devise niche-altering drugs.

Niche market

Some clinicians want to use adult stem cells to repair patients' failing bodies. Even if they don't need to invoke transdifferentiation, they may benefit from adopting an ecological mind-set. For some treatments, it may first be necessary to greatly expand stem-cell numbers by culturing them in the lab. Will the cells need to be grown in a realistic microenvironment to maintain their 'stemness', or will a simple bath of chemicals do? Some researchers argue that providing an appropriate three-dimensional environment in which signals come from the right direction will matter as much as using the right biochemicals. The same may go for getting stem cells to give rise to the appropriate tissues. "You do need an organized environment," argues Jones.

Stem cells implanted in this nude mouse created their own microenvironment and grew hair.

SALK INST.

Leanne Jones says turning stem cells into specific tissues will mean replicating their niche in the lab.

Of course, many stem-cell biologists are working with embryonic stem cells, rather than adult stem cells. Found in blastocysts — embryos consisting of a hollow ball of cells — embryonic stem cells have the potential to give rise to all of the body's tissues. They don't live in such complex microenvironments as their adult counterparts, but the new wave of ecological thinking may still help to bring embryonic stem cells into the clinic.

"Just injecting embryonic stem cells is probably not the way to go," says Fuchs. "Left to their own devices, they do their own thing and go in different directions." Instead, most researchers working with embryonic stem cells are trying to get them to differentiate into specific cell types in the lab. But to unlock the cells' potential fully, biologists may need to find ways to recapitulate the changing microenvironments that characterize the long journey from embryonic stem cell to adult tissue.

With an ecosystem as complex as the human body, attempting to recreate the precarious balances found in nature is no small task. To find the optimal conditions for stimulating stem cells therapeutically, biologists will need to know how stem cells and their neighbours depend on one another for survival. And they must test how far a system's equilibrium can be pushed before stem cells are wiped out or turn cancerous. The parallels with the local species extinctions and population explosions that can occur if a natural ecosystem is disturbed are clear.

Many stem-cell biologists have already made their travel plans for next month's meeting of the International Society for Stem Cell Research in San Francisco. Perhaps some should put another date in their diary: the Ecological Society of America meets in Montreal, Canada, from 7 to 12 August. ■

Kendall Powell is a freelance writer based in Broomfield, Colorado.

1. Wolf, N. S. &. Trentin, J. J. *J. Exp. Med.* **127,** 205–214 (1968).
2. Xie, T. & Spradling, A. C. *Science* **290,** 328–330 (2000).
3. Kiger, A. A., Jones, D. L., Schulz, C., Rogers, M. B. & Fuller, M. T. *Science* **294,** 2542–2545 (2001).
4. Kai, T. & Spradling, A. *Nature* **428,** 564–569 (2004).
5. Zhang, J. *et al. Nature* **425,** 836–841 (2003).
6. Calvi, L. M. *et al. Nature* **425,** 841–846 (2003).
7. Song, H., Stevens, C. F. & Gage, F. H. *Nature* **417,** 39–44 (2002).
8. Shen, Q. *et al. Science* **304,** 1338–1340 (2004).
9. Wurmser, A. E. *et al. Nature* **430,** 350–356 (2004).
10. Blanpain, C., Lowry, W. E., Geoghegan, A., Polak, L. & Fuchs, E. *Cell* **118,** 635–648 (2004).

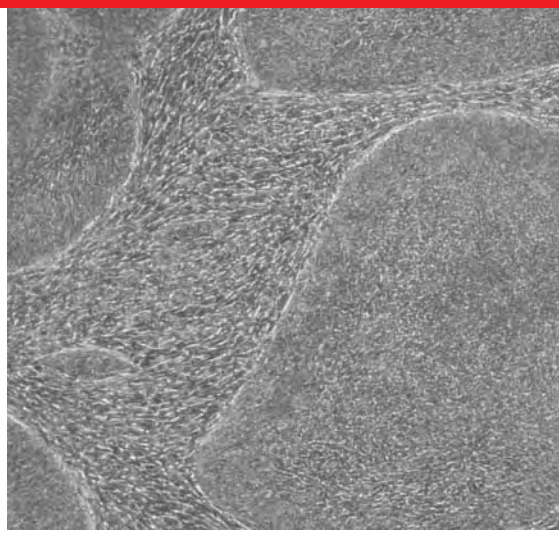
REF. 10/ELSEVIER

©2005 Nature Publishing Group

BUSINESS

Licensing fees slow advance of stem cells

Would-be entrants to the embryonic stem-cell business find access to the technology prohibitively expensive, Meredith Wadman reports.



Hard cell? Carl Gulbrandsen (top right) says that licence terms for human embryonic stem-cell lines are reasonable — but Jeanne Loring argues that they are pricing companies out of business.

R. GONZALEZ/WARF

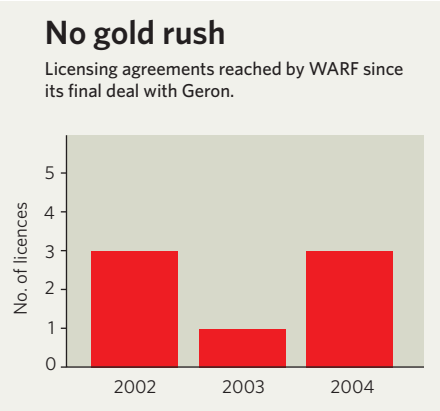
Jeanne Loring, an embryologist at the Burnham Institute in La Jolla, California, knows all about the unrealized commercial promise of human embryonic stem-cell research. A decade ago she developed a method for deriving embryonic stem cells, and in 1999 founded Arcos BioScience to culture the cells and develop cellular therapies from them. By the end of 2000, Loring had derived nine human embryonic stem-cell lines.

But that's where the success story ended. The company, already on a shoestring budget, couldn't find the $100,000 up-front fee that would allow it to take out a commercial research licence from the major patent holder in the field, the Wisconsin Alumni Research Foundation (WARF).

The foundation administers the intellectual property of the University of Wisconsin in Madison, where human embryonic stem cells were isolated by biologist James Thomson in 1998. It holds broad patents on a method of deriving human embryonic stem cells — and on the cells themselves.

But Loring says it has a stranglehold on the field. "The greatest roadblock to the development of human embryonic stem-cell research in the United States is WARF's fundamental patent," she says. "If it hadn't been for that patent, I would still be in biotech."

Instead, Loring's company merged with two others, and she went back to academic stem-cell research. By forcing researchers out of the commercial sector, she says, stem-cell patents are slowing the drive to get applications into the clinic.



Seven years after WARF wrapped up its broad package of patents in the United States — it hasn't won similar patents elsewhere — complaints are mounting about the fees and conditions it imposes on commercial firms.

"The intellectual-property situation is stifling industrial research and investment in this area," complains a senior executive at one firm that aspires to enter the embryonic stem-cell business. He requested anonymity because his company is in negotiations with WARF.

WARF says it is trying to speed the development of the field while fulfilling its mandate to support research at the University of Wisconsin — and keeping the promises it made to donors who provided the embryos from which its stem-cell lines were derived. "We've tried to be a responsible citizen," says Carl Gulbrandsen, the foundation's managing director.

Paying the price

Dozens of companies around the world are working on stem cells, exploring uses that range from drug discovery to drug screening, and from toxicology to cellular therapies for Parkinson's and heart disease. But most deal with stem cells derived from adults, which are out of the reach of WARF's patents. Those that invest seriously in embryonic stem-cell work can be counted on the fingers of two hands.

About half these firms are based outside the United States, but even these will probably want to sell any products they develop in the lucrative US healthcare market. And that means facing another major obstacle: exclusive licences held by Geron, based in Menlo Park, California, which paid for much of the Wisconsin research.

Geron has secured exclusive US commercial rights to what are generally agreed to be the three most clinically promising types of cell derived from embryonic stem cells: cardiac myocytes, which could help to repair damaged hearts; neural cells, which might treat neurodegenerative diseases or spinal-cord injury; and pancreatic β-cells, which could treat diabetes.

Partly because Geron's control is so broad, "things are moving more slowly than they otherwise would", says Rebecca Eisenberg, a law professor at the University of Michigan who specializes in intellectual-property issues in biotechnology. "The proprietary constraints are the price you pay for having to make do with private rather than public funding."

Eisenberg contrasts the situation with the handling of recombinant DNA technology in the 1980s and 1990s, when Stanford and the University of California liberally administered patents to allow free access for academics, and low-cost licensing to companies.

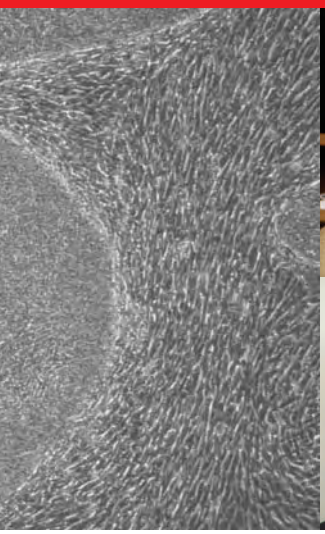
Academic scientists are also griping about WARF's charges for access to the five stem-cell lines that it owns. The foundation charges $5,000 each time it supplies one of these to an academic scientist, and doesn't let them share the cells with colleagues.

"I'm a senior investigator, and I was able to get a $5,000 grant supplement from the National Institutes of Health," says Lawrence Goldstein, a cell biologist and Howard Hughes Medical Institute investigator at the University of California, San Diego. He recently bought some cell lines from WARF, but is upset that he isn't allowed to share them with less well-funded junior colleagues: "That's not how you make science proceed, for crying out loud!"

WARF has sent cell lines to 231 research groups worldwide and signed material-transfer agreements with more than 200 US academic institutions, Gulbrandsen points out. He adds that it has lost $1.3 million by distributing the cell lines through the WiCell Research Institute, a subsidiary that WARF established in 1999.

Gulbrandsen says he is weary of getting a bum rap from all concerned. Strict rules preventing sharing of those lines between university researchers are needed, he explains, so WARF can honour promises made to the donors from whose leftover embryos the cells were generated. Each scientist or company that

©2005 Nature Publishing Group



receives a cell line signs an agreement not to do experiments to implant the line in an embryo, generate an embryo, or implant an embryo in a uterus. If WARF allowed free sharing, it couldn't police this, Gulbrandsen notes.

He doesn't deny that WARF is seeking six-figure sums — a $25,000 annual maintenance fee plus $100,000 up-front fee — from companies taking out commercial licences. "But we've tried to be accommodating where a smaller company thinks that the price is too high," he says — in one case, WARF accepted equity in a firm instead of cash.

WARF has issued seven commercial licences to companies (see chart, left) since its first broad patent in 1998, but declines to identify them. Becton, Dickinson & Company, a New Jersey-based drugmaker, says it has a licence. Advanced Cell Technology (ACT) of Worcester, Massachusetts, told *Nature* that it, too, has signed a deal with WARF. *The Wall Street Journal* reported last month that General Electric and Novartis are about to launch US projects with embryonic stem cells — General Electric will develop drug-testing products for sale to pharmaceutical companies, and Novartis aims to turn stem cells into heart cells. And Johnson & Johnson, the New Jersey-based maker of medical products, has bought an equity stake in Novocell of Carlsbad, California, which seeks to generate insulin-secreting pancreas cells from stem cells.

The disgruntled blame the small number of licencees on WARF's prices and restrictions. "It granted only seven commercial licences in seven years, on a technology that is hot — why?" demands the anonymous executive. He adds that US-based firms feel hamstrung by one requirement in particular: "If we sign an agreement with WARF in the United States, it wants to place restrictions on us globally, even though the patents do not apply worldwide."

But Robert Lanza, vice-president of medical and scientific development at ACT, says that the price his company paid the foundation for a commercial research licence was "very fair". In WARF's position, many companies might overcharge licensees, he says. "And WARF isn't. My hat's off to them." ■

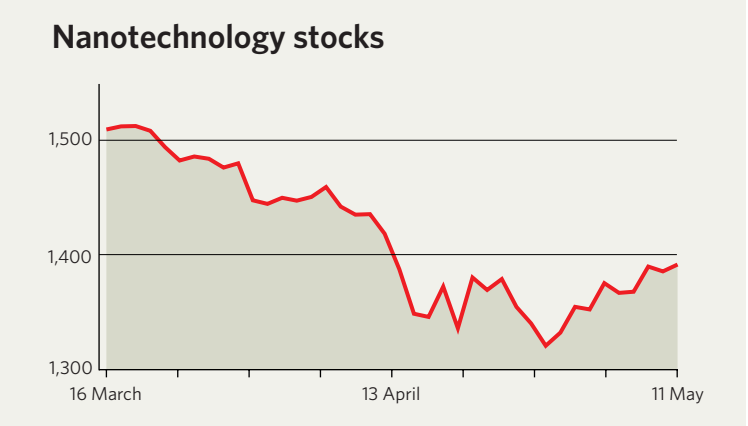
IN BRIEF

DISEASE TEST APPROVED The US Food and Drug Administration has approved the first genetic blood test for cystic fibrosis. The Tag-It test, made by TM Bioscience of Toronto, Canada, can now be used to identify children and adults who carry the disease. Cystic fibrosis is the most common inherited fatal disorder, afflicting about 1 in 3,000 babies in the United States. The test identifies only some of the 1,300 genetic variations associated with the disease, and the terms of the approval require that it be used with other approaches to diagnose cystic fibrosis.

VACCINE MAKERS FLOUNDER The number of US companies manufacturing vaccines has fallen from 26 in 1967 and 17 in 1980 to just 5 last year, and shows no sign of reviving. Writing in this month's *Health Affairs*, Paul Offit, head of infectious-disease research at the Children's Hospital of Philadelphia, says that high development costs, low revenues and the threat of legal liability are forcing manufacturers out of the vaccine business. Offit says the US government should provide more financial incentives for vaccine development and amend a 1986 law designed to limit manufacturers' exposure to legal liability for ill-effects caused by vaccines.

UP IN SMOKE European car makers are falling short of voluntary targets agreed with the European Commission to increase fuel efficiency and decrease carbon dioxide emissions from their vehicles. Average new-vehicle emissions fell by 1.8% last year, the *Financial Times* reports, against a 3.3% average reduction needed to meet the emissions target of 140 g per km by 2008. European manufacturers have launched smaller models, such as the BMW 1 Series, with a view to improving the figures. The industry fears that the European Union will set compulsory limits if the voluntary one isn't met.

MARKET WATCH



SOURCE: LUX RESEARCH

Nanotechnology stocks have lost ground rapidly over the past two months, according to one of the first stock-market indices devoted to tracking them. But analysts still expect the field to attract $400 million in venture capital this year — the most it has managed since 2002.

The downturn reflects the general doldrums in technology shares this year, says Peter Hebert, chief executive and co-founder of Lux Research, the consultancy in New York that developed the index.

"Most of it has to do with the overall market for technology stocks," he says, adding that 'the Lux' has taken an even harder hit this year than the Nasdaq — the main US index for technology stocks — because the former is weighted towards smaller firms, whose shares tend to suffer most when investors are feeling cautious.

But 2005 is shaping up to be a bumper year for venture-capital investment in nanotechnology. Lux reports that $66 million was raised in March alone for three US companies — Nanomix, Nantero and Nano-Tex — and estimates that the flow of venture capital into US nanotechnology firms will double this year from the $200 million raised during 2004.

Tracking the financial performance of something as loosely defined as nanotech isn't easy, Hebert concedes. Lux has had a stab by pulling together companies that supply nanotech products, build nanotech tools such as atomic-force microscopes, or use nanotech in their businesses. Its index allocates equal weighting to all of the firms — effectively giving more clout to the smaller, more specialized ones.

"We've been one of the first to stress that nanotechnology isn't an industrial sector per se," says Hebert. "It's an enabling technology. Investors have been perplexed about how to approach nanotech — but we're seeing progress now." ■

www.luxresearchinc.com

©2005 Nature Publishing Group

When science meets religion in the classroom

SIR – In the Editorial "Dealing with design" (*Nature* **434,** 1053; 2005), *Nature* claims that scientists have not dealt effectively with the threat to evolutionary biology posed by 'intelligent design' (ID) creationism. Rather than ignoring, dismissing or attacking ID, scientists should, the editors suggest, learn how religious people can come to terms with science, and teach these methods of accommodation in the classroom. The goal of science education should thus be "to point to options other than ID for reconciling science and belief". In this way, students' faith will not be challenged by scientific truth, and evolution will triumph.

This suggestion is misguided: the science classroom is the wrong place to teach students how to reconcile science and religion. For one thing, many scientists deem such a reconciliation impossible because faith and science are two mutually exclusive ways of looking at the world. For such scientists, *Nature* apparently prescribes hypocrisy. The real business of science teachers is to teach science, not to help students shore up world-views that crumble when they learn science. And ID creationism is not science, despite the editors' suggestion that ID "tries to use scientific methods to find evidence of God in nature". Rather, advocates of ID pretend to use scientific methods to support their religious preconceptions. It has no more place in the biology classroom than geocentrism has in the astronomy curriculum.

Scientists are of course free (some would say duty-bound) to fight ID outside the classroom, or to harmonize religion with science. But students who cannot handle scientific challenges to their faith should seek guidance from a theologian, not a scientist. Scientists should never have to apologize for teaching science.

Jerry Coyne
Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA

Peter Atkins *Lincoln College, University of Oxford*
Colin Blakemore *Medical Research Council, London*
Richard Dawkins *Oxford University Museum, University of Oxford*
Steve Jones *Galton Laboratories, University College London*
Richard Lewontin *Museum of Comparative Zoology, Harvard University*
John Maddox *9 Pitt Street, London W8 4NX*
Paul Nurse *The Rockefeller University, New York*
Linda Partridge *Department of Biology, University College London*
James D. Watson *Cold Spring Harbor Laboratories, New York*
Steven Weinberg *Department of Physics, University of Texas, Austin*
Lewis Wolpert *Department of Anatomy and Developmental Biology, University College London*

Teaching about ID helps students see its flaws

SIR – I have regularly taught seminars for university biology majors, which compare the scientific claims of evolution and ID. In doing so, I am not advocating the scientific merits of ID, as discussed in your News Feature "Who has designs on your students' minds?" (*Nature* **434,** 1062–1065; 2005). I view these seminars as analogous to media literacy courses. To understand why 80% of Fox News viewers had misperceptions about Iraq, such as believing that weapons of mass destruction had been discovered there (see www.pipa.org), media students need to learn how Fox News operates. Such a media literacy course does not necessarily vouch for the veracity of any particular Fox show.

My interest in ID was sparked in 1999 by a local high-school teacher who used ID materials in a biology course. Parents and citizens successfully defended the teaching of mainstream science against proponents of ID, in this case the Discovery Institute (see www.scienceormyth.org). This taught me how effective pro-evolution groups are when they work with the school administration, and are supported by faculty from local colleges and universities. But to be effective in its support, the scientific community needs to understand the empirical claims of ID.

Although it seems to have been resurrected for religious or cultural agendas, ID's proponents have made empirical claims that can be examined. Many college students are curious about ID but have little knowledge of the claims made for it. In my experience, upper-level biology students with the appropriate background in molecular biology, genetics, developmental biology and evolution are capable of distinguishing the scientific merits of evolutionist and ID claims — to the great disadvantage of ID.

Students who themselves determine that ID does not cut the scientific mustard will be more effective in their support of teaching mainstream science. Students who remain creationists or fence-sitters will at least have a better understanding of why ID has not been widely accepted in the scientific community.

It may seem contradictory to offer a course on ID and evolution in colleges and oppose teaching ID in high schools. But high-school students are just learning the basics of science. To expect them to make a well-reasoned judgement about the status of any scientific theory, including evolution, is unrealistic.

David Leaf
Department of Biology, Western Washington University, Bellingham, Washington 98225, USA

Evolution is a short-order cook, not a watchmaker

SIR – The visual shock of last week's 'intelligent design' cover was matched by the perceptiveness of your News Feature (*Nature* **434,** 1062–1065; 2005) on the seepage of this slyly religious ideology into science curricula. This stuff should certainly be kept out of high schools, but I am ambivalent about its presence in our universities, where free discussion is of greater value than correctness — political, scientific or otherwise.

On the other hand, I'd be properly rebuked and sanctioned for incompetence if I were to assert in my undergraduate physical-chemistry course that the intricate, precisely exponential distribution of velocities observed in a collection of gas molecules is simply too perfect and beautiful to have arisen from random collisions, but that we should instead consider a mechanism by which intelligent designers — let's call them "Maxwell's angels" — individually push on each molecule, while keeping in intelligent communication with each other to maintain this distribution.

One reason that scientists famously fail in rebutting ID is that we use the wrong analogies. Evolution is not a blind watch-maker or any other kind of engineer, but rather a short-order cook, and — looking at the phenomenally complicated structures — one who is less like Isaac Newton than Rube Goldberg or W. Heath Robinson.

A terrific argument against ID came to me recently after two consecutive talks, one on the Wnt signalling pathway, the next on G-protein crosstalk in control of cellular calcium. Just look at the details, and you'll immediately abandon all thoughts that biological systems were designed with any intelligence whatsoever.

Chris Miller
HHMI, Department of Biochemistry, Brandeis University, Waltham, Massachusetts 02454, USA

Seeking evidence of God's work undermines faith

SIR – Your Editorial about the promotion of ID in schools and universities (*Nature* **434,** 1053; 2005) asks us to persuade our students that science and faith do not compete, but for Christians this should always have been clear. In the Bible (John 20: 25–29), Thomas doubts that the man speaking to him is the resurrected Christ until Jesus reveals his wounds. Thomas then believes, but Jesus says: "Blessed are those who have not seen and yet have believed".

The Bible throughout teaches that faith is

©2005 Nature Publishing Group

more valuable when expressed in the absence of evidence. For a Christian, when science is allowed to be neutral on the subject of God, science can only bolster faith. In contrast, and I imagine without realizing it, ID proponents have become professional Doubting Thomases, funded by Doubting Thomas Institutes. When advocates of ID use the vocabulary of science to argue for God's presence in cellular machinery or in the fossil record, they too poke their fingers through Jesus' hands. In so doing, ID vitiates faith.

Not realizing this, many Christians now believe they are making a stand against evil by supporting religion-infused alternatives to evolution. For them, the fundamental debate is not over which is wrong and which is right, but over which is good and which is bad, and the majority opinion is clear. So if we want to ensure the continued learning of evolution in our schools, we cannot only argue that science and faith can be reconciled; we also have to show that ID actively undermines the basis of Christianity.

Douglas W. Yu
Centre for Ecology, Evolution and Conservation, School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ, UK

Leave well alone and stick to teaching what you know

SIR – Your Editorial "Dealing with design" (*Nature* **434,** 1053; 2005) is another piece of evidence of the peculiar angst among certain scientists about the ID strategy of a rather robust fraction of the US population. On the basis of some decades of work in this area, I do not believe that your advice to those who feel so threatened is wise, for two reasons.

There are some very skilled experts on the topic of how to deal with different cultures or belief systems. Their advice, from experience, would be: leave well alone. Act like a scientist, confident in your own — always tentative, always open to change — axioms and laws. Read the literature, for God's (or Darwin's) sake. It will prove to you that even graduates of MIT and Harvard do not know simple scientific facts that are irrelevant to their work, such as why the Earth experiences winter and summer, despite having been explicitly taught such facts several times during their education. This amazing ignorance does not affect their performance as scientists. I do not know a single materials scientist or engineer whose technical work would be affected by their beliefs about evolution/ID. My advice: relax. It can do very little harm. Ham-fisted efforts will simply alienate much larger numbers of people from the rest of science.

As to the suggestion that scientists should "offer some constructive thoughts of their own": beware of the ignorance, nay illiteracy, of many scientists on matters of social and political concern. I recommend Huston Smith's book *Why Religion Matters* (HarperSanFrancisco, 2002) for advice on how to handle the ID debate.

"Building a straw man based on natural selection alone makes it easy for opponents to poke holes in evolution" — Michael Lynch

Rustum Roy
The Pennsylvania State University, 102 MRL, University Park, Pennsylvania 16802, USA

Intelligent design or intellectual laziness?

SIR – Much of the concern over ID (*Nature* **434,** 1053 and 1062–1065; 2005) has focused on veiled attempts to inject religion into public education. Sheltered within the confines of academia, most biologists find it hard to believe that the slain need to be slain again. Those in the trenches — school boards, school biology teachers and their national representatives — often don't know how to respond, in part because they themselves never really achieved a deep understanding of evolutionary biology at college.

However, there is a related and equally disturbing issue: the legitimization of intellectual laziness. Have a problem explaining something? Forget about it: the Designer made it that way. Any place for diversity of opinion as to who/what the Designer is/was? The ID literature makes it very clear that there is no room for scientific discourse on that. Think I'm exaggerating? To get a good idea of what IDers would have the face of science look like, check out the journal *Perspectives on Science and Christian Faith* (www.asa3.org/ASA/PSCF.html).

Two factors have facilitated the promotion of ID. First, IDers like to portray evolution as being built entirely on an edifice of darwinian natural selection. This caricature of evolutionary biology is not too surprising. Most molecular, cell and developmental biologists subscribe to the same creed, as do many popular science writers. However, it has long been known that purely selective arguments are inadequate to explain many aspects of biological diversity. Building a straw man based on natural selection alone makes it easy for opponents to poke holes in evolution. But features of the genome, such as genomic parasites or non-coding introns, which aren't so evolutionarily favourable (nor obviously 'intelligent' innovations), can be more readily explained by models that include random genetic drift and mutation as substantial evolutionary forces.

Second, IDers like to portray evolution as a mere theory. But after a century of close scrutiny, evolutionary theory has passed so many litmus tests of validation that evolution is as much a fact as respiration and digestion.

Less widely appreciated is that evolution has long been the most quantitative field of biology, well grounded in the general principles of transmission genetics. Yet few students at university, and almost none at high school, are exposed to the mathematical underpinnings of evolutionary theory. The teaching of evolution purely as history, with little consideration given to the underlying mechanisms, reinforces the false view that evolution is one of the softer areas of science.

Here is a missed opportunity. Our failure to provide students with the mathematical skills necessary to compete in a technical world is a major concern in the United States. Mathematics becomes more digestible, and even attractive, when students see its immediate application. What better place to start than with the population-genetic theory of evolution, much of which is couched in algebraic terms accessible to school students?

Michael Lynch
Department of Biology, Indiana University, Bloomington, Indiana 47405, USA

Solidarity with the oppressed flat-Earthers

SIR – I was disturbed by your News Feature "Who has designs on your students' minds?" (*Nature* **434,** 1062–1065; 2005), in which the proponents of ID are mostly portrayed as a persecuted minority. They are said to be afraid to reveal their identity and to be frequently censured into silence by anti-democratic scientists and administrators.

Your reporter clearly does not realize that 'intelligent designers' are not the only minority bullied into submission by the scientific establishment. The vast majority of flat-Earthers, tea-leaf readers, astrologers, geocentrists and phlogiston theorists cannot publish their studies in respectable journals. It is rumoured that *Nature* has rejected without review a study showing that storks bring babies into the world. I have even heard of a physician who was fired from a university hospital for trying to cure his patients by altering the ratio of blood to yellow bile and phlegm to black bile.

Thanks to your News Feature, I am now convinced that by replacing "small, medium and large" with "tall, grande and venti" — as in my local coffee-shop — the disreputable theory of biblical creationism can be turned into a respectable scientific discipline called 'intelligent design'.

Dan Graur
Department of Biology and Biochemistry, University of Houston, Houston, Texas 77204-5001, USA

©2005 Nature Publishing Group

BOOKS & ARTS

Reason to fight back

A stout defence of the evidence-based approach from attacks on all sides.

The March of Unreason: Science, Democracy, and the New Fundamentalism
by Dick Taverne
Oxford University Press: 2005. 320 pp. £18.99, $29.95

John Durant

The barbarians are at the gate — and what a motley crew they are! Alongside religious fundamentalists of all sorts there are purveyors of alternative medicines, organic farmers, eco-warriors and even a few postmodernist philosophers and sociologists, who have had the temerity to call for wider public consultation over science and technology policy. What unites all of these apparently disparate people, according to Dick Taverne, is their disregard, or even disdain, for the "evidence-based approach" of science, and their wish to elevate "unreason" (in the form of various kinds of dogma, mysticism or personal prejudice) above reason in the conduct of human affairs.

The March of Unreason is a bracing affirmation of Enlightenment values against any and all nay-sayers. In a semi-autobiographical prologue, Taverne recounts his early conversion to the cause of environmentalism in the 1960s, and his subsequent disillusionment as reasonable and pragmatic concerns for the welfare of the environment steadily lost out to more extreme and ideologically driven forms of 'eco-fundamentalism'. Green warriors, he tells us, have fostered public suspicion about science and mistrust of experts, to the point where scientists have come to blame themselves for public ignorance of, or misgivings about, their work. All this, Taverne argues, constitutes a direct threat to both science and democracy, because "the scientific method and democracy are natural allies and unreason is their common enemy".

There is much to agree with and even to admire in Taverne's wide-ranging and trenchant observations. Certainly, contemporary society is now less straightforwardly optimistic about, and deferential towards, science and technology than it was in the past. But for all the apparently more questioning and critical attitudes, our culture also indulges a surprising amount of pure hokum. For example, pharmaceutical companies are required to spend millions of dollars testing new drugs, but practitioners of alternative medicine are free to purvey all manner of herbs, potions and other nostrums (including pure water!) that have never been properly tested.

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

P. PAVANI/AFP/GETTY IMAGES

The needs of the developing world may be trampled underfoot by protesters against transgenic crops.

Before happily concluding that alternative treatments generally make great placebos and may thus potentially benefit many patients, we would do well to recall Taverne's warnings: choosing ineffective alternative therapies over effective conventional ones can be dangerous; and further, many alternative therapies are worse than ineffective as they have various (often poorly characterized) clinical effects that can cause direct harm to those who take them. There should not be one rule for conventional medicine and another for alternative medicine; all candidate treatments should be tested as rigorously as possible for safety and effectiveness.

Similar comments can be made about much that Taverne has to say about food and farming. In recent years, conventional farming has come in for growing criticism, and the flood of (generally unsubstantiated) claims for the virtues of organic farming has risen to a virtual torrent. A key ingredient in the considerable success of the organic-food lobby has been the vilification of genetically modified (GM) food. It is true that early applications of GM technology in Western Europe and North America offered few serious benefits to consumers (who in any case enjoy an abundance of food and food choices in the supermarket), but the emergence of what Taverne would call fundamentalist opposition to GM technology in agriculture could do serious harm to the prospects of many people in the developing world.

In the late 1990s I chaired a series of public debates about GM food across Britain. I watched as public opinion turned inexorably away from GM food and towards organic farming. The only time I saw an audience pause and begin to move the other way was when an academic plant scientist described his public-domain research using GM technology to create new root-worm-resistant varieties of staple crops that might constitute a lifeline for marginal farming communities in the Caribbean. Relatively rich environmentalists in the developed world will have a great deal on their consciences if their lobbying efforts against GM food impede or prevent vitally important biotechnological research that might otherwise have provided real help to farmers in poorer parts of the world.

In these and other ways, then, *The March of Unreason* is to be applauded. Nevertheless, the book has some real weaknesses. In the main, these come from trying to squeeze too many different issues into the strait jacket of "reason versus unreason". It is one thing, perhaps, to criticize various kinds of religious fundamentalism as enemies of the "evidence-based approach". But when religious fundamentalists are lumped together with radical environmentalists, animal-rights activists, homeopathic medical practitioners, organic

©2005 Nature Publishing Group

farmers, anti-globalization campaigners and various kinds of academic philosopher and sociologist, one wonders whether the epithet "unreason" has lost its critical edge. What do all these different groups really have in common? Are they on the same side, intellectually or practically speaking? And if so, what side is it?

The problem is at its most severe where Taverne deals with intellectual criticism of various types of scientific and science-policy practice. Noting the call for "more democratic science", he concedes that the public and its representatives have an important role to play in the development of science. But (in an argumentative style that is repeated throughout the book), having conceded this important point, he immediately undermines it by blaming those who are working to make science more responsive to public interests and concerns for having "driven scientists onto the defensive". This charge immediately reduces what is an important area of constructive debate, about the way that science policy should be conducted in advanced democracies, to an 'us versus them' or (even worse) a 'reason versus unreason' stand-off.

Those of us who seriously advocate closer public engagement in science and science policy-making are not motivated by anti-scientific or antirational sentiments. Rather, we recognize that making decisions about how to conduct and apply science and technology in advanced industrial societies is a complex and difficult business. Experts of various sorts have essential roles to play, and so too do democratic representatives. But it is increasingly becoming clear that the establishment of sustainable policies in socially sensitive areas of science and technology is facilitated by the engagement of others in the process — such as special-interest groups, stakeholder groups and citizens' groups.

Frankly, tarring efforts to achieve wider engagement in science and technology policy-making with the broad brush of 'antiscience' or 'unreason' is simply not helpful. ■

John Durant is chief executive of At-Bristol, Harbourside, Bristol BS1 5DB, UK.

At the trowel's edge

The Goddess and the Bull: Çatalhöyük: An Archaeological Journey to the Dawn of Civilization
by Michael Balter
Free Press: 2005. 416 pp. $27, £18.99

Pat Shipman

This book is about neither a goddess nor a bull, unless Michael Balter is using a metaphor too subtle for me to appreciate. Indeed, *The Goddess and the Bull* is not really about the archaeological site of Çatalhöyük either. After much thought, I believe this is actually a post-processual book about archaeology.

Post-processualism is a concept developed by Ian Hodder, a Cambridge-trained archaeologist who now works at Çatalhöyük in Turkey. In its early formulation, Hodder suggested that the best way to approach archaeology is "characterized by debate and uncertainty about fundamental issues that may have been rarely questioned before". He added that archaeologists "move backwards and forwards between theory and data, trying to fit or accommodate one to the other in a clear and rigorous fashion, on the one hand being sensitive to the particularity of the data and on the other hand being critical about assumptions and theories." Post-processual archaeology is a dialogue, not a diatribe.

So too is this book, which provides a great deal of information: about archaeological theory, methodology and traditional interpretations of Çatalhöyük, one of the famous (or infamous) sites that inspired the concept of the Neolithic revolution. There have been decades of excavation by the original site director, James Mellaart, and by more than 100 specialists under Hodder's modern (or post-modern) direction. The most recent excavation, which began in 1993, is Hodder's brave attempt to integrate his theoretical stance with field practice. The new methodology includes ongoing and constantly changing "interpretation at the trowel's edge", with computer diaries written by the excavators, video recording of discussions about interpretation and methodology carried out in the trenches, and constant interactions among scientists, locals, politicians, goddess-worshippers, carpet scholars and other groups who claim some 'ownership' of Çatalhöyük's past. It is fair to say that Hodder's task has been difficult and complex.

Balter raises many compelling questions about the differing and changing interpretations of Çatalhöyük and its remarkable finds, but does not answer them. He explains the significance of Çatalhöyük as the earliest known site with domesticated cattle (a conclusion contradicted by recent zooarchaeological analyses). Although Çatalhöyük was an unusually dense settlement with bull's horns embedded in some walls, and some female figurines, no credible hypothesis is ever offered for the meaning of these odd features. In fact, the reader never learns much about the life of those who lived there, despite the astonishing number of human skeletons buried under the floors of houses. All this extraordinary evidence begs for an explanation.

There are also myriad questions about Mellaart. In the bizarre 'Dorak affair', Mellaart was purportedly shown a treasure trove of looted artefacts from northern Turkey by a mysterious woman who subsequently disappeared. Only Mellaart's drawings and descriptions of the artefacts remained. An inquiry conducted by the British Institute at Ankara exonerated Mellaart from any involvement in looting, but even so, Mellaart's excavation was shut down in 1965.

In 1989, Mellaart, together with carpet specialists Belkis Balpinar and Udo Hirsch, published *The Goddess from Anatolia* (Eskenazi), which included stunning reconstructions of 44 wall paintings from Çatalhöyük. Why had there been no word of such glorious art before? Blatant discrepancies between the book's claims and Mellaart's earlier pronouncements cast doubt on the paintings' very existence. Early reports described only red and black paint, not the striking blue in the new reconstructions. Rooms identified in the book as having magnificent wall paintings had been earlier declared by Mellaart to have no paintings. No excavator remembered seeing the fragments upon which Mellaart's reconstructions were based. All corroborative evidence had been destroyed by fire in 1967, Mellaart told Balter. As Carl Lamberg-Karlovsky remarked: "Bluntly put, there is no objective reason to believe that these 'new' wall paintings exist." Further, Mellaart proposed that

J. & A. MELLAART

The horns of a dilemma: what is the meaning of this painting of a red bull found at Çatalhöyük?

©2005 Nature Publishing Group

the paintings were linked to patterns found on Turkish kilims today, but the Çatalhöyük patterns cannot be made into rugs using the weaving technology preserved at the site.

The book details much debate but few conclusions. The result is a good read that bespeaks the importance of this enigmatic and iconic site and highlights Balter's considerable journalistic skills. The book is both accessible and fascinating. Balter tries, with moderate success, to show us how personality, nationality and the training of the scientists involved influences their scientific ideas.

Yet the book left me distinctly dissatisfied: I learned more about the childhoods of the excavation team members than about ancient Çatalhöyük. This is an intelligent, provocative book by a distinguished science writer who visited the site every field season for six years, interviewed the excavators, and read their publications, which are referenced in extensive notes and a lengthy bibliography. The scholars who have worked at Çatalhöyük are impressive, the duration of excavations far in excess of normal expectations. Why then is so much about Çatalhöyük so unclear?

Perhaps the reason is Balter's adherence to a Hodder-like reluctance to settle on a single interpretation for a site that means so much to so many. What are we to think, then, of Çatalhöyük and its evidence, excavators, myths? That remains the post-processual question. ■

Pat Shipman is in the Department of Anthropology, Pennsylvania State University, 315 Carpenter Building, University Park, Pennsylvania 16802, USA.

Hidden depths

Fathoming the Ocean: The Discovery and Exploration of the Deep Sea
by Helen M. Rozwadowski
Belknap: 2005. 304 pp. $25.95, £16.95, €24

The Remarkable Life of William Beebe: Explorer and Naturalist
by Carol Grant Gould
Shearwater: 2004. 416 pp. $30

Descent: The Heroic Discovery of the Abyss
by Brad Matsen
Pantheon: 2005. 286 pp. $25

Jon Copley

Deep-sea science is big science. Ocean covers 365 million square kilometres, and most of it is more than two kilometres deep. To understand what goes on down there, you need a ship to brave the high seas and equipment that can reach into the abyss. As today's researchers agonize over grant proposals and publication records, some may yearn for the time when they could chart the depths without worrying about tenure or research assessment exercises.

WILDLIFE CONSERVATION SOCIETY

The descent of man: William Beebe (left) and Otis Barton used their bathysphere to explore the ocean depths.

But as these three books charting the history of deep-sea science reveal, that golden age never existed.

Fathoming the Ocean by Helen Rozwadowski chronicles the birth of deep-sea oceanography, from early observations by Benjamin Franklin to the voyage of HMS *Challenger* in the 1870s. She weaves a rich narrative from the work of renowned as well as lesser-known oceanographers. While unearthing the foundations of the subject, she reveals some striking parallels with modern research careers.

Like today, there was plenty of job-hopping, with worries about money and research output. When Edward Forbes accepted a chair in botany at King's College, London, in 1843, he also became curator of the museum at the Geological Society of London to boost his income. But he was concerned that he no longer had any time for research, and jumped ship just a year later for a job with the Geological Survey. This strategic jockeying paid off, and he was later appointed regius chair in natural history at the University of Edinburgh.

Then there is the tale of George Wallich, who sailed as a naturalist on the cable-surveying voyage of HMS *Bulldog*. Wallich hoped the expedition would make his name in scientific circles, as other voyages of discovery had done for T. H. Huxley and Darwin. But it was not to be. Despite initial enthusiasm about his results, Wallich failed to secure election to the Royal Society. Under the financial pressures of supporting his wife and children, he became a photographer instead. He described the prospects of his new career as "more than I could venture to hope for in that muddy sea of science". His story may sound familiar to today's postdocs-turned-plumbers.

Worrying about funding also occupied the mind of deep-sea pioneer William Beebe. To write *The Remarkable Life of William Beebe*, Carol Gould was granted unprecedented access to Beebe's personal papers that he had bequeathed to his colleague Jocelyn Crane. Grant's detailed and well organized biography is a treasure. From the waters of Bermuda to the jungles of Venezuela, Beebe was tireless in his enthusiasm for understanding the living world, and he provided the inspiration for many scientific careers.

Brad Matsen's *Descent* focuses on Beebe's collaboration with Otis Barton and their bathysphere dives. In the 1930s, they plunged six times deeper than anyone before and became the first people to see deep-sea life *in situ*. "No human eye had glimpsed this part of the planet before us," wrote Barton, often considered the more prosaic of the pair, "this pitch-black country lighted only by the pale gleam of an occasional spiralling shrimp." Matsen offers a worthy tribute to their remarkable achievement, and explores the tensions between them. His account is captivating, although not as lavishly referenced as Gould's biography.

In the days before research councils and national science foundations, Beebe was using publicity and popular accounts of his work to charm funds from philanthropists. Like some who popularize their research today, he sometimes encountered snobbery from his academic peers. But deep-sea research has always been newsworthy and captured the public imagination. On 26 April 1857, the front page of *The New York Herald* hailed the laying of the first transatlantic cable as the "great work of the age", and illustrated the story with microscope drawings of seafloor sediments. Seventy-eight years later, radio listeners right across the United States and Western Europe tuned in to hear Beebe's voice live from the bathysphere at a depth of

NATURE EDITOR WINS AVENTIS BOOK PRIZE

Philip Ball, a science writer and consultant editor of *Nature*, has won this year's Aventis Prizes for Science Books General Prize. *Critical Mass: How One Thing Leads to Another* (William Heinemann) takes a look at the application of physics to the collective behaviour of society. Bill Bryson, who chaired this year's judging panel and won the prize in 2004, says: "This is a wide-ranging and dazzlingly informed book about the science of interactions. I can promise you'll be amazed." (For a review of this book see *Nature* **428,** 367–368; 2004.)

Robert Winston takes the junior Aventis Prize for his children's book *What Makes Me, Me?* (Dorling Kindersley). The judging panel for this prize included schoolchildren as well as adult writers and scientists.

The winners received their awards at a ceremony on 12 May 2005 at the Royal Society in London.

©2005 Nature Publishing Group

670 metres. Beebe vividly described his visit to another world, three decades before the televised Moon landings.

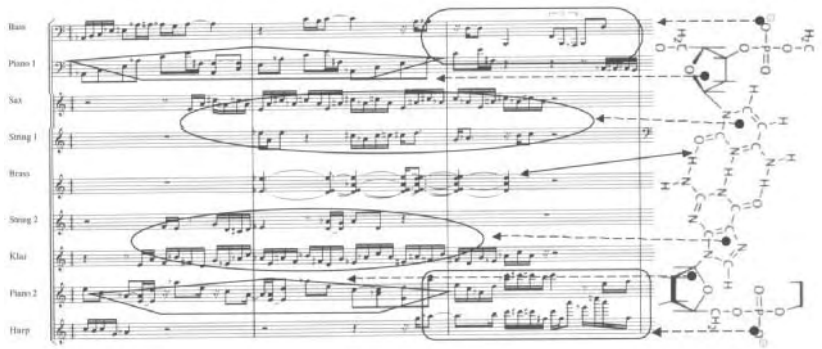
Most modern grant proposals require applicants to describe the wider benefits of their work to society. Those studying the deep sea can point to examples of medical treatments, industrial enzymes and even tips for making better optical fibres based on the glass skeletons of deep-sea sponges. Rozwadowski describes how early workers highlighted the benefits of seafloor dredging for cable surveys when lobbying for the use of HMS *Lightning*, HMS *Porcupine* and HMS *Challenger*. But Gould and Matsen show that Beebe got funding by promising a payback in the joy of knowledge itself. At the age of 16, he wrote that "to be a Naturalist is better than to be a King". Taken together, these books reveal how far we have come in understanding the largest habitat on our planet — and how much further we have to go. ■

Jon Copley is at the National Oceanography Centre, Southampton SO14 3ZH, UK.

SCIENCE IN CULTURE

The music of life

Composer Thilo Krigar seeks to represent the flow of genetic information.



THILO KRIGAR

Juliane Mössinger

Music and science seem irreconcilable to many, but there is a close and long-standing relationship between them. Since antiquity, music has been categorized as *scientia mathematica*, together with arithmetic, geometry and astronomy. In the eighteenth century, Johann Sebastian Bach composed some almost mathematically constructed canons as part of his *Musical Offering*. And Karlheinz Stockhausen used the Fibonacci series to determine the durations of the 33 Moments in *Mikrophonie II* in 1965. More recently, the sciences have offered patterns in data, such as those arising from genomics, as inspirations for musical composition.

But previous attempts to portray DNA, the icon of twentieth-century science, in musical terms have been limited to the computerized conversion of nucleotide sequences into audible patterns. The composer Thilo Krigar, however, is not interested in simply representing the structure or sequence of the DNA molecule itself, but in a musical exploration of the flow of genetic information. He is fascinated by the idea that the dynamic and continuous renewal of our cells relies on the information coded in our DNA. He perceives this process of 're-creation' as the ultimate form of creativity — a metaphor for composition.

The Berlin-based cellist and composer has been working on *DNA in Concert* for the past five years, using the atoms that compose the DNA molecule as a starting point. He converts the number of outer valence or total electrons of hydrogen, carbon, nitrogen, oxygen and phosphorus into an equivalent number of semitone steps. These then form five different musical intervals, which, in turn, are the basis for the melodic and harmonic structure of the composition.

Krigar then uses other musical tools to represent the biochemistry of the cell. Stability in the harmonic architecture of the music, expressed by a double octave, for example, represents the chemical stability of hydrogen bonding. Although Krigar is interested in a close musical portrayal of biochemical processes, he realizes that a representation of all of the atoms, molecules and chemical bonds is virtually impossible and, musically, may not be desirable. Above all, he wants the resulting music to be aesthetically appealing and demanding.

DNA in Concert begins with a prelude that depicts the archaic process of reverse transcription — the mechanism thought to be involved in the transition from RNA into DNA on the early Earth. The main body of the work, 'Seasons4life', is then divided into four parts: Transcription, Translation, Metabolism and Replication.

During Transcription, the music reflects the physical tension in the DNA caused when RNA polymerase, the enzyme that copies DNA into RNA, unwinds and rewinds the template DNA helix. Saxophones represent the three-dimensional structure of the double helix, and the strings the actual generation of RNA molecules.

The music of Translation is focused on the activity of the ribosomes, complex biochemical machines that catalyse the translation of a nucleic-acid message into a protein sequence. Interjections of percussion, for example, represent amino acids as a building material. Live electronic sound embodies the energy consumed during protein assembly.

In Metabolism, the chemical signals that connect the activities of the cellular proteins are transformed into musical signals that are passed on from one musician to another. Different groups of instruments compete with one another through intensifying motifs to illustrate the polymerization of the two DNA daughter strands during replication.

Newly developed harmonies then portray the processes involved in cell division in the concluding passage, Proliferation, before the music retreats back to the structure of DNA.

The music is modern but not atonal. It is reminiscent of minimalism but has elements that resemble the dramatic romanticism of a Mahler symphony. The piece is written for a chamber ensemble, and the musicians are expected to move around the room, giving the audience the feeling of being surrounded by the double helix and the biochemical processes in the cell. Loudspeakers projecting computer-generated sound add to the surround effect. Digital movies of cellular processes bring a visual component to the performance.

The composition is not limited to repeating sequential tones, nor does it simply attempt a one-to-one representation of the molecular reality in the cell. The composer's work is based instead on biological concepts, stressing the parallels between biological processes and human creativity. He wants to give the listener a musical experience of the biochemistry in the cell and to become progressively aware of the life process taking place within ourselves.

DNA in Concert is a work in progress. Its world premiere — minus an unfinished coda — will be performed by the Pythagoras Strings on 28 and 29 May 2005 at TESLA im Podewils'schen Palais in Berlin, Germany.

Juliane Mössinger is an associate editor in the physical sciences at *Nature*.

www.dna-in-concert.de

©2005 Nature Publishing Group

Capturing chaos

Ergodicity: a fundamental assumption of statistical physics — anything that can happen will happen — was thrown into question 50 years ago. Now it looks solid after all.

Mark Buchanan

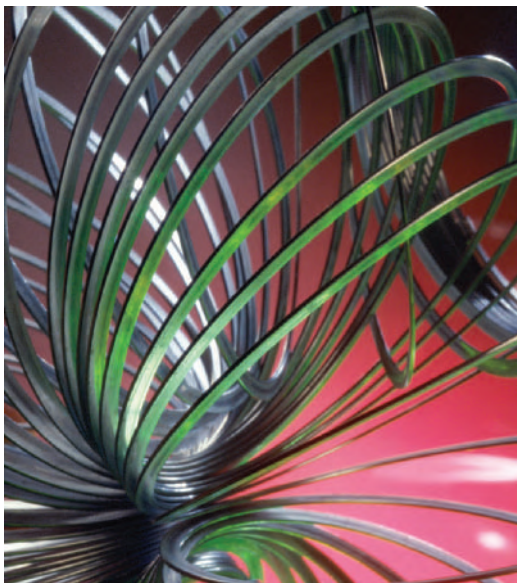
In 1954, at the Los Alamos National Laboratory in New Mexico, Enrico Fermi, John Pasta and Stanislas Ulam undertook an experiment that was unusual for the time. They used one of the early computers — a hulking device with thousands of vacuum tubes called MANIAC I — to simulate the dynamics of a long chain of masses linked together by springs. Setting their virtual chain in motion with an organized, wave-like vibration, they sought to measure how quickly that motion would degenerate into random chaos; how the chain, started off out of equilibrium, would relax into it.

"The results of our computations," Fermi and colleagues later reported, "were, from the beginning, surprising." The organization did not degenerate, it persisted. For as long as they could run their computer, the chain never settled into equilibrium.

Fermi and his colleagues' experiment cast doubt on a fundamental assumption of physics — the so-called 'hypothesis of ergodicity' — and thereby on the foundations of the physics of solids, liquids and other forms of matter. Fifty years of further work have finally vindicated those foundations, and may offer a penetrating new perspective on some very old problems.

The notion of ergodicity asserts, in a sense, that anything that can happen will happen; that a system having a number of possible states will, over a finite time, visit each and every one with equal frequency. A fly might spend all day in one corner of a room, or it may buzz around but never visit the window at the far end. Neither behaviour would be ergodic. An ergodic fly would go everywhere, exploring every last part of the room repeatedly and spending, in the long run, the same time in each area.

The assumption of ergodicity is a kind of democratic principle of dynamics. In physics, it offers a way to build up knowledge on a basis of ignorance. In the Fermi–Pasta–Ulam chain — as in any bit of ordinary matter — nonlinear interactions between too many particles make it impossible to know exactly how the system will evolve. Moreover, the system might well spend more time in some states than in others, making it equally impossible to get any general understanding, even of its average behaviour, in the long run. Knowing next to nothing, however, one might assume, boldly, that all states get equal treatment.



GETTY IMAGES

Springing back: a chain can be forced to take up all possible states, so long as it is given a big enough kick.

This assumption wipes away a world of more complicated possibilities. And, as a result, a theorist only has to average over all of a system's states, without bias. This idea works beautifully in practice, in literally thousands of cases. Without it there would be no theory of liquids or conductors or magnets.

But Fermi and his colleagues' experiment made all this success look like a miraculous and unwarranted gift, for their simple chain did not explore all its states equally, but got hung up, returning repeatedly to specific wave-like patterns of vibration. The chain violated all the rules of ergodicity — hence 50 years of continued interest with the experiment. Only recently, by using much faster computers, have physicists got to the bottom of the conflict.

To set their chain in motion, Fermi and colleagues gave it some energy with an initial kick. In seminal work over the past decade, researchers have explored systematically how the behaviour of such a chain changes with increasing energy, with illuminating results. As it turns out, ergodicity seems to come into play when the energy given to the chain is about ten times greater than that applied by Fermi and colleagues in their original study. At this energy, rather than remaining locked into some kind of semi-repetitive state, the vibrating chain begins to explore its possible states ergodically and relaxes slowly into equilibrium.

Importantly, this relaxation happens more quickly as the number of linked masses increases; longer chains get hung up less easily. Similar behaviour has been found in several other simple systems, suggesting that something like ergodicity typically reigns when the number of particles involved is very large — as is the case in ordinary matter. Statistical physics, it seems, has been saved.

These studies have also discovered a second transition that occurs at higher energy — from so-called 'weak' chaos to 'strong' chaos. This switch seems to be intimately linked to abrupt phase transitions wherein matter turns from one organized form into another. Until now, phase transitions have been understood in statistical terms, with little detailed connection made to the underlying microscopic dynamics. But this dynamic change suggests that phase transitions may be understood in another way, as reflecting an abrupt qualitative transformation in the way a system explores its possible states.

It is ironic that a century ago the foundations of statistical physics were taken to be sound, and to rest firmly on the ergodic hypothesis. It took the insight of Fermi and his colleagues, and the fastest computer of the time, to suggest that there might be a problem. Fifty years, a lot more thinking, and immeasurably faster computers were then needed to show that things were fine after all. So physicists have come back to where they started, but it all looks very different. ■

Mark Buchanan is a science writer based in Cambridge, UK.

FURTHER READING

Fermi, E., Pasta, J. & Ulam, S. *Los Alamos Sci. Lab. Rep. LA-1940* (1955).
Pettini, M., Casetti, L., Cerruti-Sola, M., Franzosi, R. & Cohen, E. G. D. *Chaos* **15,** 015106 (2005).
Casetti, L., Pettini, M. & Cohen, E. G. D. *Phys. Rep.* **337,** 237–341 (2000).
Lieberman, M. A. & Lichtenberg, A. J. *Regular and Stochastic Motion* (Springer, New York, 1983).

CONCEPTS

©2005 Nature Publishing Group

NEWS & VIEWS

B. JANSSON/ALAMY; M. EVERTON/CORBIS; C. MCLENNAN/ALAMY

EVOLUTIONARY BIOLOGY

Geography and skin colour

Jared Diamond

Human skin comes in many different shades. Recent studies of geographical differences in skin colour open up the subject scientifically by offering sophisticated accounts of the basis of this variation.

The most obvious — and most discussed — aspect of human geographical variability is skin colour. Most people would say that skin colour becomes darker towards the Equator to give more protection against tropical sunlight. But that claimed correlation of skin colour with latitude is riddled with exceptions, and that functional interpretation of the correlation is debated. Most scientists shy away from the whole subject because it so interests racists, and the motives of scientists studying it become suspect.

Jablonski and Chaplin[1–3] have now brought order to this confused field, starting with quantitative measurements of skin colour and sunlight. By convincingly identifying the strongest correlate of skin colour, they open the door for anthropologists to explore other correlates and exceptions.

Skin colour was formerly described qualitatively by matching it against coloured tablets, but Jablonski and Chaplin tabulate numerical values, obtained by skin reflectance spectrophotometry. And instead of using latitude as a proxy for sunlight, Jablonski and Chaplin tabulate ultraviolet radiation (UVR) itself at the Earth's surface. UVR does decrease with latitude, because at high latitudes the oblique angle at which sunlight falls on the atmosphere results in a longer atmospheric path, and hence more absorption and scattering of UVR. But the correlation of UVR with latitude is imperfect: UVR also increases with altitude owing to atmospheric thinning (for example, UVR is high on the Tibetan and Andean altiplanos); it also decreases with atmospheric water vapour in the form of rain, clouds or humidity (UVR is higher in the Atacama Desert, southwestern United States, and the Horn of Africa, than in adjacent, wetter areas to the west or east).

In this quantitative database, variation in UVR proves to be the strongest predictor of skin reflectance, explaining 77% (Northern Hemisphere) or 70% (Southern Hemisphere) of its variation. The causes of this correlation have been the subject of many theories, such as protection against skin cancer, protection against overproduction of vitamin D, and camouflage in tropical jungles.

Jablonski and Chaplin prefer a combination of two selective factors involving several costs and one benefit of UVR. The costs involve the destructive photolysis of many compounds, of which Jablonski and Chaplin attach particular importance to the B vitamin folate. Everybody requires folate, so everybody would have dark skins (to screen out UVR and reduce photolysis) if there were no other selective factors. However, UVR also provides a benefit: catalysing the synthesis of vitamin D. Hence skin colour evolves as a compromise between skins light enough to permit UVR penetration for vitamin D synthesis, but dark enough to reduce folate photolysis.

This compromise results in paler skins, to permit vitamin D synthesis, at high latitudes with low UVR levels. Albino schoolchildren in South Africa need less dietary vitamin D than normally pigmented children to attain target serum levels of calcium and vitamin D. Women in all populations tend to have paler skins than men, presumably through natural selection owing to their greater need for calcium and vitamin D during pregnancy and lactation. This difference may be reinforced by sexual selection.

Although the harmful effects of UVR are often taken to be especially associated with skin cancer, Jablonski and Chaplin minimize the importance of this aspect on the grounds of the late age of onset, after some or most of a couple's children have been born. Most geneticists similarly minimize the selective importance of damage or disease late in life.

Here, I am sceptical. This common view may be valid for mammal species in which parent–offspring ties are severed soon after weaning, but it overlooks three of the most distinctive features of human life-histories in traditional societies: the long post-weaning dependence of offspring on parents for learning and then for social status; the large contribution of hunter-gatherer grandmothers to their grandchildren's food supply[4]; and the dependence of an entire band or village on its oldest people as the repositories of knowledge in a preliterate society[5]. It would be interesting to try to estimate the selective importance of skin cancers for skin-colour evolution from this perspective: data on skin-cancer incidence (for example of Europeans in tropical

©2005 Nature Publishing Group

NATURE|Vol 435|19 May 2005

Australia) could be combined with estimates of the three contributions of 'post-reproductive' adults to the status and survival of their children and grandchildren.

Now that the strongest correlation of skin colour (with UVR) has been established, we can look for the exceptions, and the other factors involved. Jablonski and Chaplin have made this easy, by providing a table (Table 6 on pages 74–75 of ref. 3) listing residuals between observed skin reflectances and those predicted by UVR alone, for 85 indigenous human populations representing all continents. I predict that this table will be one of the most frequently cited features of their paper for anthropologists, geneticists, historians and linguists. The following correlates of tribal histories and lifestyles leap out from even a cursory examination of these residuals.

The fourth largest negative residual of skin reflectance (skin is darker than expected from UVR considerations alone) is for a Greenland Inuit population. These people 'can get away with unexpectedly dark skins' because they traditionally obtained their vitamin D from a diet rich in marine mammals. As a result, modern Inuit, subsisting instead on supermarket food, suffer from one of the world's most severe incidences of vitamin D deficiency. Conversely, three of the nine most positive residuals (skins unexpectedly pale) are for populations from the interior of Asia, remote from any seafood.

Of the 12 most negative residuals (skins unexpectedly dark), the highest and four others are for Bantu-speaking southern African populations that migrated from the Equator towards high southern latitudes only 2,000 years ago. These populations have not yet had enough time for evolutionary loss of their equatorially adapted dark skins. Conversely, three of the nine most positive residuals (skins unexpectedly pale) are for peoples of the Philippines, Vietnam and Cambodia, who migrated towards the Equator from high latitudes only in recent millennia and have not yet evolved appropriately dark skins. But why do the people of Bougainville Island in the South Pacific, and why did the Aboriginal Tasmanians, have such dark skins, after more than 10,000 years of *in situ* adaptation?

As Jablonski and Chaplin point out, skin is by far the most extensive visible feature of our bodies, a signal advertising age, health and ancestry, and a poster board for decoration. I hope that their papers will convince more scientists not to be ashamed of being interested in skin colour. ■

Jared Diamond is in the Department of Geography, 1255 Bunche Hall, University of California, Los Angeles, California 90095-1524, USA.
e-mail: jdiamond@geog.ucla.edu

1. Jablonski, N. G. *Annu. Rev. Anthropol.* **33,** 585–623 (2004).
2. Chaplin, G. *Am. J. Phys. Anthropol.* **125,** 292–302 (2004).
3. Jablonski, N. G. & Chaplin, G. *J. Hum. Evol.* **39,** 57–106 (2000).
4. Hawkes, K. *et al. Am. J. Hum. Biol.* **15,** 380–400 (2003).
5. Diamond, J. *Nature* **410,** 521 (2001).

EARTHQUAKES

Future shock in California

Duncan Agnew

For California, probabilistic principles can be applied to the short-term forecasting of further ground-shaking following an earthquake. How such predictions will be used by the public remains to be seen.

"What can we expect next?" — this question is foremost in the minds of the public and news media following any widely felt earthquake. Put another way, for any person living near the earthquake the question is, "What does this earthquake mean for the earthquake risk in my locality?". The seismologist's usual reply is a vague comment that because all earthquakes are followed by smaller events (aftershocks), this one will be too, and that although earthquakes are sometimes followed shortly afterwards by larger ones, this is rare, so that the overall risk is essentially unchanged, except close to the epicentre.

Thanks to the work of Gerstenberger *et al.* (page 328 of this issue)[1], a much more precise answer can be given, at least for California. The procedures they describe also set a new standard against which to test future earthquake predictions, and as such they deserve to be adopted, with appropriate local modifications, in other earthquake-prone areas.

The simplest kind of earthquake forecast assumes randomness, with the probability of a particular-sized earthquake occurring in a particular place over (say) the next day being always the same. This is the Poisson model. Such a forecast also assumes (as is observed) that small earthquakes occur much more often than big ones, a rule that, when expressed more precisely, is known as the Gutenberg–Richter law. This simple forecast also needs to include a description of how the probabilities vary from place to place; this can be controversial away from the boundaries of tectonic plates, but in California can be reasonably estimated from what we know about past earthquakes and geologically active faults.

Gerstenberger *et al.*[1] add to this a sophisticated model for the other, well-established behaviour of earthquakes — that big events are followed by smaller aftershocks, at a rate that decreases with time. This is known as the Omori law, after the Japanese scientist who suggested it in 1895. An important element in Gerstenberger and colleagues' paper is to allow aftershocks to be of any size, including bigger than the original earthquake, using the Gutenberg–Richter law again to describe how likely the different sizes of aftershocks are[2]. This neatly unifies the common case of aftershocks with the much less common case of the first earthquake becoming, after a larger second one, a foreshock. The authors use a range of aftershock models to match the variability in observed aftershocks (Fig. 1); as the

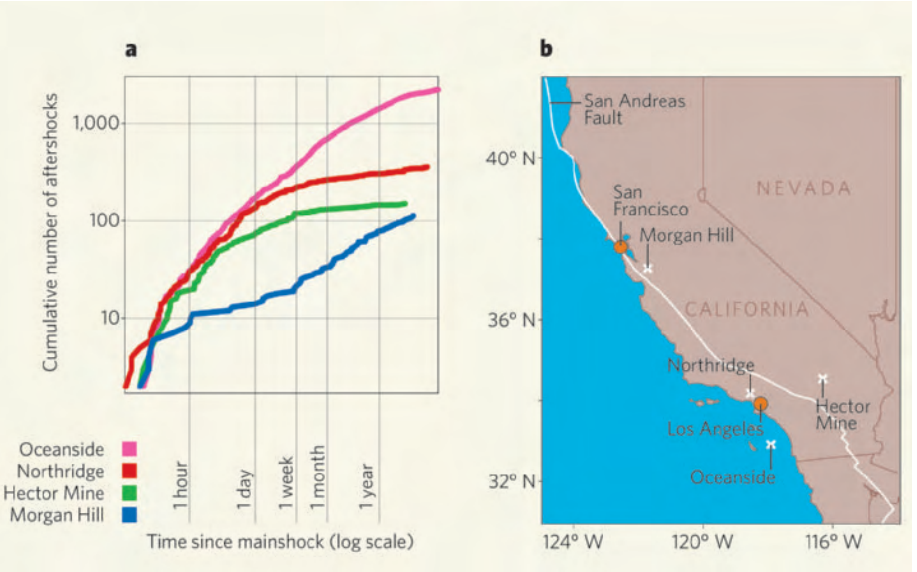


Figure 1 | Variability in aftershocks for four earthquakes in California. a, The cumulative numbers of aftershocks (within 3.5 magnitude units of the mainshock) against the log of elapsed time. **b**, The locations of the earthquakes, which are Oceanside (magnitude 5.4; July 1986), Northridge (6.6; January 1994), Hector Mine (7.1; October 1999) and Morgan Hill (6.2; April 1984). Gerstenberger and colleagues' forecasting procedure[1] automatically allows for the wide variability seen in the numbers and decay rates of aftershocks.

©2005 Nature Publishing Group

aftershocks occur, the increasing amounts of data accumulated can be used to improve the prediction of future behaviour.

The result is a procedure that can be run in real time to predict the probabilities of different sizes of earthquakes, with long-term expectations being modified by recent events. The final step is to put this into a form that answers the question about local risk. For any particular location in California, the procedure first computes the probability of earthquakes occurring throughout the region, and finds the shaking that all of these other hypothetical earthquakes would produce. Combining these produces a map of the probability of shaking at a certain level over some future time interval, in this case the next 24 hours. For maximum usefulness to the public, the authors use a shaking level at what seismologists call intensity VI (causing cracked plaster and broken windows), because this is the lowest level with significant associated costs.

To demonstrate their method, they apply it retrospectively to the 1992 Landers earthquake (see Fig. 2 of the paper on page 329). Just before the main earthquake, the local probability of shaking in the area is somewhat increased because of aftershocks from an earlier event. Just after the quake, the probability becomes much larger (up to about a 10% chance of shaking the next day), diminishing over time as the aftershocks become less frequent.

This kind of short-term earthquake prediction has two values, one for public understanding and policy, the other scientific. Having such information readily available (as it will be, on the web) provides both reassurance about what is happening and a clearer picture of the risks. The public is accustomed to seeing probability numbers given for weather forecasts, and seismologists are fortunate in now being able to provide them for earthquakes — avoiding the vagueness that has dogged the colour-coded terror alert system in the United States. It remains to be seen in what ways, besides their educational value, such short-term (and generally low) probabilities will be used by the public and the emergency services.

The scientific value of these predictions is that they provide a baseline against which to test other predictive schemes. Progress in earthquake prediction depends on moving beyond single, anecdotal accounts to systematic tests of predictions using statistical methods. Such a procedure, pioneered by the late Frank Evison in New Zealand[3,4], has now become common elsewhere[5,6], and is explicitly deployed in Gerstenberger and colleagues' paper. They show that their model does significantly better than the Poisson model, which is the usual comparison[7].

It is not surprising that a model that includes aftershocks does better than one that does not. But now that we have it, it can be taken as the hypothesis against which other predictive schemes should be tested, and challenged to demonstrate better performance. I hope the methodology can be extended to other earthquake-prone regions, so that its benefits, and the challenge it presents, can be applied there as well. ■

Duncan Agnew is at the Institute of Geophysics and Planetary Physics, Room 221, 8800 Biological Grade, La Jolla, California 92037, USA.
e-mail: dagnew@ucsd.edu

1. Gerstenberger, M. C., Wiemer, S., Jones, L. M. & Reasenberg, P. A. *Nature* **435,** 328–331 (2005).
2. Reasenberg, P. & Jones, L. M. *Science* **243,** 1173–1176 (1989).
3. Evison, F. F. & Rhoades, D. A. *NZ J. Geol. Geophys.* **36,** 51–60 (1993).
4. Evison, F. F. & Rhoades, D. A. *NZ J. Geol. Geophys.* **40,** 537–547 (1997).
5. Kagan, Y. Y. & Jackson, D. D. *Geophys. J. Int.* **143,** 438–453 (2000).
6. Schorlemmer, D. *et al. J. Geophys. Res.* **109,** B12308 doi:10.1029/2004JB003235 (2004).
7. Stark, P. B. *Geophys. J. Int.* **131,** 495–499 (1997).

MOLECULAR MOTORS

Kinesin steps back

Justin E. Molloy and Stephan Schmitz

Kinesin is a protein motor that ferries membrane-bound packages around cells — but only in one direction. Forcing it into reverse provides clues to its inner workings and to how molecular machines might be engineered.

Kinesin can carry a packet of neurotransmitter from your spine to the tip of your finger in about two days — a journey that would take a thousand years if left to simple diffusion. This molecular motor 'steps' along the microtubules that form the cytoskeletal scaffolding of cells, by converting chemical energy into mechanical work. Carter and Cross (page 308 of this issue)[1] probe the stepping behaviour of kinesin by using optical tweezers to push and pull on a single kinesin molecule attached to a small plastic bead. When they pulled strongly enough, kinesin could be made to walk backwards in a sustained manner. This dramatic result could provide the basis for developing a motor to drive nanoscale molecular machines.

There are 14 families of closely related kinesin proteins, and conventional kinesin-1 is a two-headed motor that moves along microtubules towards the so-called 'plus' end. The coordinated activity of its two motor heads ensures that at least one of them remains tightly bound to the microtubule at all times. It therefore takes hundreds of processive steps before falling off its microtubule track. The 8-nm steps match the repeat distance of the microtubule lattice[2], and we have a fair insight into the molecular mechanisms underlying them[3]. For instance, we know that each step requires the breakdown (hydrolysis) of one molecule of adenosine triphosphate (ATP, the energy currency of the cell), that the two kinesin heads move in 'hand-over-hand' fashion[4], and that movement stalls when the backward load exceeds about 7 piconewtons.

Carter and Cross[1] address two new points. First, how does each kinesin head move forwards to its next binding site on the microtubule — is it in a single rapid jump, or a series of more complex 'sub-steps' powered by articulations of the molecule? Second, can the kinesin motor be made to step backwards along the microtubule? If so, then it would give insights into coordination between the two motor heads.

By observing the time course of each step, the authors looked for clues about the molecular event that drives kinesin to its next binding site. If kinesin unbinds and moves in one hop by diffusion (the 'thermal ratchet' mechanism), then close scrutiny of individual steps should show no defined pauses. However, if movement is driven by a conformational change in the kinesin head, or 'power stroke', then there might be consistent breaks or sub-steps within the movement. In fact, Carter and Cross did not observe sub-steps, so kinesin movements must be completed within the (excellent) time resolution of their measurements (less than 50 μs). Although the data do not argue strongly for either mechanism, they do set important limits on our modelling of the kinesin stepping motion.

By applying high loads to the kinesin, the authors examined whether kinesin can be made to move backwards, rather than simply being ripped from the microtubule. They found that by suddenly applying a super-stall force, kinesin would indeed walk backwards along the microtubule. It will be interesting to know whether, when it walks backwards, all the corresponding biochemical and mechanical steps are exactly reversed in sequence so that kinesin resynthesizes ATP from its hydrolysis products ADP and inorganic phosphate (P_i). We know that the F_OF_1 ATP synthase complex, which is a rotary motor, works like this. This complex couples ATP synthesis or hydrolysis with the flow of protons across the mitochondrial membrane. When it spins in one direction[5], it hydrolyses ATP; but during its normal operation[6], it is driven in the opposite direction by the proton flux and manufactures ATP from ADP and P_i.

Experience tells us that only certain motors

©2005 Nature Publishing Group

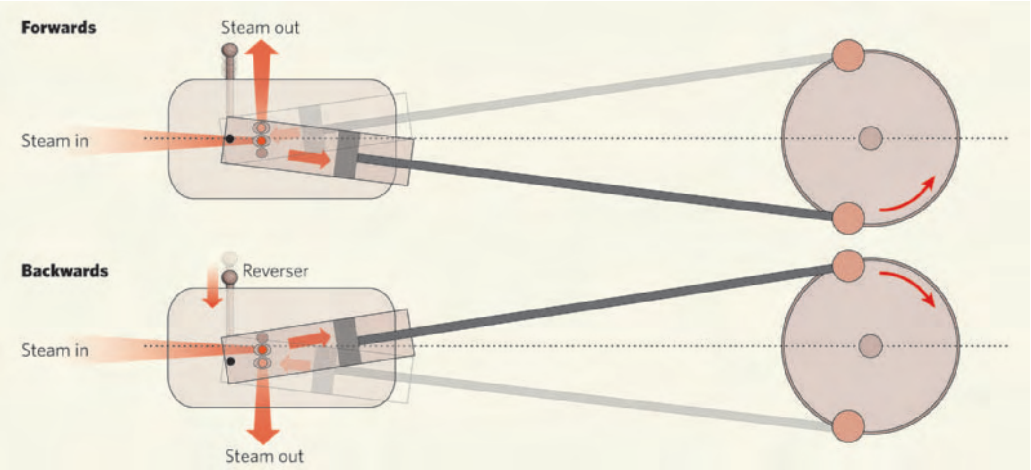


Figure 1 | A motor in reverse. A toy steam engine can be made to run forwards or backwards at will, by changing the position of the cylinder mounting to adjust its valve mechanism. Carter and Cross discovered[1] that when conventional kinesin is suddenly subjected to a very high load, it will often stagger backwards in a process fuelled by ATP. It probably does this by changing the phasing of the ATPase cycles of its two 'heads', similar to the valve changes in the toy.

can act as generators when driven in reverse. For instance, a small electric motor can act as a dynamo, but steam engines clearly cannot resynthesize coal when driven backwards. However, if valve timing is unchanged and the engine is pushed backwards, air is pumped back into the boiler, so energy in the form of pressurized gas is generated instead. Significantly, if valve timing is shifted (by half a cycle) and the steam engine is pushed backwards, steam is now allowed out and no generator action occurs.

Carter and Cross[1] report that this sort of thing seems to happen with kinesin, whereby phasing of the ATP hydrolysis cycles on the lead and trailing heads seems to be swapped during backwards movement. The front head detaches first and then reattaches behind the old trailing head, and so on. So if we consider the catalytic sites to behave as valves, then when the motor is pulled forcibly back, the timing of this valve operation seems to be perturbed. This means that fuel (ATP) and products (ADP and P_i) still enter and leave in the correct order to allow ATPase activity to proceed, but the two heads actually step backwards. If true, kinesin would use the same amount of ATP fuel when it walks forwards as it does staggering backwards under load.

This brings to mind toy steam engines that can be flicked into reverse by shifting the position of the cylinder mounting (Fig. 1). This changes the valve timing relative to piston position and hints at a way of engineering kinesin to run backwards by somehow tinkering with its 'valve' timing. Perhaps by altering the linkage between the kinesin head and the rest of the molecule, the way in which the load acts upon the catalytic site could be suitably rearranged[7] (see Fig. 4 on page 311).

Recently, another motor protein called myosin was made to work backwards[8] by a slightly different strategy. The rotational sense of the output lever arm of myosin was reversed by folding its light-chain binding domain back upon itself, redirecting its output. This is similar to the idea of engaging the reverse gear in a car — the engine still rotates normally, but its output transmission direction is reversed by the gearbox.

From a basic scientific standpoint, interesting physics, chemistry, biology and maths are all wrapped up in kinesin. Using genetic engineering combined with various single-molecule assays, we can now test our ideas directly to see whether we can make motors run in different ways. So, perhaps we are not very far from making a completely synthetic nanomachine, and perhaps such a machine will rival the impact of the steam engine. ■

Justin E. Molloy and Stephan Schmitz are in the Division of Physical Biochemistry, MRC National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK.
e-mail: jmolloy@nimr.mrc.ac.uk

1. Carter, N. J. & Cross, R. A. *Nature* **435,** 308–312 (2005).
2. Svoboda, K., Schmidt, C. F., Schnapp, B. J. & Block, S. M. *Nature* **365,** 721–727 (1993).
3. Knight, A. E. & Molloy, J. E. *Nature Cell Biol.* **1,** E87–E89 (1999).
4. Yildiz, A., Tomishige, M., Vale, R. D. & Selvin, P. R. *Science* **303,** 676–678 (2004).
5. Noji, H., Yasuda, R., Yoshida, M. & Kinosita, K. *Nature* **386,** 299–302 (1997).
6. Abrahams, J. P., Leslie, A. G., Lutter, R. & Walker, J. E. *Nature* **370,** 621–628 (1994).
7. Henningsen, U. & Schliwa, M. *Nature* **389,** 93–96 (1997).
8. Tsiavaliaris, G., Fujita-Becker, S. & Manstein, D. J. *Nature* **427,** 558–561 (2004).

PARTICLE PHYSICS

Vanishing pentaquarks

Frank Close

After a first inconclusive sighting, the search for exotic particles that consist of five quarks has been hotly pursued in the past few years. But the weight of evidence is now shifting against their existence.

Correct perceptions differ from mistaken ones in that they become clearer when experimental accuracy is improved — Irving Langmuir's observation may have gained a new example with the latest report on the question of the 'pentaquark' particle. A high-statistics experiment at the Jefferson Laboratory in Virginia finds no evidence to support claims of the existence of this enigmatic object that have been made over the past three years. Final tests of the data remain to be completed, and a second independent experiment is still in progress. But this is a serious setback for the many who had hoped that this novel particle had revealed unexpected phenomena in quantum chromodynamics (QCD), the theory governing the 'strong' interactions of quarks — subatomic particles thought to be elemental and indivisible.

The story began in 1997 with the prediction[1] that an analogue of the proton should exist with a mass of "about 1,530 MeV" (some 50% more massive than the proton) and with both positive electric charge and a positive value for another fundamental quantum number, strangeness.

Such a correlation was not possible within the simplest quark model, where most strongly interacting particles (known as hadrons) are either mesons, which contain a quark and an antiquark, or baryons, which comprise three quarks. To make such a correlation of charge and strangeness requires a particle consisting of four quarks and one antiquark (hence dubbed a 'pentaquark'). Such combinations are allowed by QCD but are expected to be highly unstable, with 'widths' of many hundreds of MeV. (An inherent property of quantum mechanics is that short lifetimes are correlated with large uncertainties in energy. Thus the mass, or energy at rest, of a short-lived particle is actually a distribution with an intrinsic width — for strongly interacting unstable particles, such widths

©2005 Nature Publishing Group

NATURE|Vol 435|19 May 2005

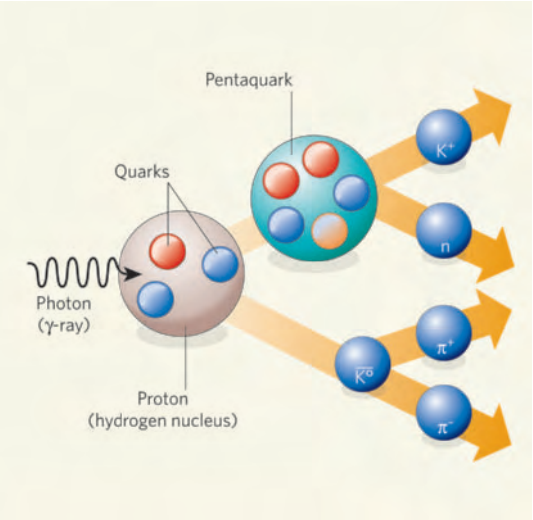


Figure 1 | Give me five. Representation of the reaction channel investigated by the CLAS experiment at the Jefferson Laboratory for evidence of a pentaquark state. An incident photon (γ-ray) initially interacts with a hydrogen nucleus, consisting of one proton. The main constituents of a proton are three 'valence quarks' of different types, or 'flavours': two 'up' (blue circles) and one 'down' (red). At sufficiently high interaction energies, the conservation laws of particle physics would allow a short-lived pentaquark (five-quark) resonance to be created — containing, in addition to up and down quarks, a 'strange' antiquark (yellow circle). This pentaquark does not live long enough to be observed directly; rather, its fleeting existence must be inferred from longer-lived products of this particular reaction — a K meson (also with strange-quark content), a neutron and two π mesons. From its reconstruction of such reactions, CLAS has failed to find any evidence for the existence of a pentaquark.

typically exceed 100 MeV.) No pentaquark had ever been seen with certainty, and their absence had been one of the planks upon which the standard quark model had been developed.

A surprising feature of the 1997 prediction was that the particle would have a width of the order of a few MeV and not the hundreds that might have been expected. Initially, the paper[1] received little attention, but the LEPS collaboration at the SPring-8 laboratory in Japan was encouraged to mount an experiment to look for the particle. The first pentaquark sighting was announced by them[2] in early 2003, exactly at the predicted mass and with a narrow width.

The experiment involved photon beams interacting with protons or neutrons in a carbon nucleus. There was some surprise that the first sighting of a particle with such a narrow width should have occurred in such a complex environment: the nuclear constituents are bound and have kinetic energy, which tends to smear any signals. However, this stimulated experimentalists elsewhere to look again at their data from earlier experiments to see if these contained evidence for the pentaquark.

Within a few months, teams from the Jefferson Lab, from Russia and from the SAPHIR collaboration at the Electron Stretcher Accelerator (ELSA) in Bonn, Germany, all announced that they, too, had spotted tantalizing hints of the particle in data taken in other experiments. For instance, the SAPHIR team's evidence of the pentaquark[3] came from data they had obtained in 1997–98 and confirmed its mass of 1,540 MeV. None of these experiments on their own was very significant, but the broad agreement among them created huge excitement.

By the end of 2003, more than ten experiments worldwide had reported evidence for the pentaquark (see ref. 4 and references therein), mostly produced by photons interacting with protons or neutrons ('photoproduction'). The pentaquark particle then decayed into a K meson and a proton or neutron. The targets included protons, and deuterium and heavier nuclei; the kinematics covered both low and high energy; and the narrow peak invariably occurred around 1,540 MeV. Hints of sibling pentaquarks also emerged, for example one with a positive value of another quantum number, charm, rather than strangeness. Well over 1,000 theoretical papers have addressed such phenomena.

In 2004 a series of theoretical criticisms emerged, centred around some anomalies. On closer inspection there seemed to be small but systematic differences in the mass, and in the width of the signals[4]; also, it was unclear how such a narrow width state was apparently produced so readily. Moreover, reports of negative experimental searches began to appear. The null results tended to come from experiments using nuclei, or hadrons such as π mesons or protons, rather than photons, and also included searches involving very-high-energy electron beams. Unlike some of the supposedly positive sightings, the common feature of the null results was that they tended to have rather large statistical samples.

One suggestion was that the pentaquark might have some unusual production mechanism, such that photoproduction at energies of a few GeV is especially favoured. This loophole could be closed by dedicated photoproduction experiments with high statistics, and two such experiments had already been designed at the Jefferson Lab.

The latest news from researchers in the Jefferson Lab's Large Acceptance Spectrometer (CLAS) collaboration, announced at the American Physical Society spring meeting on 16 April, adds to the concern about the reality of the pentaquark. The researchers have taken data from an experiment in which a photon beam hit a liquid-hydrogen target (Fig. 1), under conditions similar to those of the earlier experiment conducted by the SAPHIR collaboration. The CLAS team's data contain statistics that are improved by two orders of magnitude, and find no evidence of a pentaquark with mass 1,540 MeV.

The CLAS collaboration data show, at a level of precision at least 50 times higher than the published SAPHIR result, that this particular reaction produces no pentaquark. Researchers at the Jefferson Lab are currently undertaking dedicated hunts for the pentaquark, including an experiment that repeats their original pentaquark search with much higher statistics. Those data are being analysed, and the results are expected later this year. If they show a null result, the pentaquark story will probably have come to an end for physicists but will live on as a case-history for historians and philosophers of science. ■

Frank Close is in the Rudolf Peierls Centre for Theoretical Physics, University of Oxford, 1 Keble Road, Oxford OX1 3NP, UK.
e-mail: f.close@physics.ox.ac.uk

1. Diakonov, D. *et al.* preprint at www.arxiv.org/hep-ph/9703373 (1997).
2. Nakano, T. *et al.* (LEPS collaboration) *Phys. Rev Lett.* **91,** 012002 (2003).
3. Barth, J. *et al.* (SAPHIR collaboration) preprint at www.arxiv.org/hep-ex/0307083 (2003).
4. Zhao, Q. & Close, F. E. *J. Phys. G* **31,** L1–L5 (2005).

NEUROSCIENCE

Plasticity and its limits

Martin I. Sereno

How much can the adult brain compensate for injury to the senses of touch or vision, for example? The answer from the latest results on the visual system, involving damage to the retina, seems to be 'very little'.

Many sensory systems are characterized by connections from receptor surfaces, such as the retina or skin, in which the relationship between neighbouring inputs is preserved. The resulting 'topological maps' are also commonly maintained in subsequent projections between sensory areas within the brain. Early work showed that these map-like projections can adapt to the loss of input from part of the receptor sheet by rearranging input lines, with the silenced brain areas regaining responsiveness to remaining parts of the sheet.

In the somatosensory system, which among other things conveys the sensation of touch, this plasticity was unexpectedly found to persist into adulthood. Now, however, Smirnakis *et al.* (page 300 of this issue)[1] show that in the initial stages of the primate visual

©2005 Nature Publishing Group

system topological map-like projections are not plastic in the adult on a timescale of half a year.

The ability of the developing brain to adapt to damage is well documented. A striking example in humans is that children who have had their left cerebral hemisphere removed early in life often regain motor control of the right side of their bodies, and go on to develop almost normal language abilities using the right hemisphere[2,3]. As the brain matures, however, it becomes much less plastic. For example, even though adults may become proficient in a second language, they find it difficult to erase all traces of a foreign accent. Children, by contrast, can learn the fine vocal distinctions of a second language with native-speaker precision.

Given the reduction in plasticity with age, Merzenich and colleagues' demonstration in 1983, that the primary somatosensory cortex in monkeys remained plastic in the adult, came as a shock[4]. It had long been known that sensory maps in some vertebrates — such as that from the retina in frogs — remained plastic into adulthood[5]. But this was thought to be due to the fact that, in the frog, both the retina and the target brain area concerned — the optic tectum — add neurons throughout adult life.

In the experiment of Merzenich *et al.*, damage to nerves carrying touch information from several fingers in monkeys initially resulted in silencing of most of the corresponding finger maps in primary somatosensory cortical area 3b. After several weeks, however, the silenced region came to be activated by skin from the surrounding fingers — the representations of those fingers expanded to completely fill in the silent cortex. Although the body-map reconfigurations in the early experiments were measured in millimetres, much longer-term denervation of an entire arm resulted in reconfigurations of more than one centimetre[6] — a distance likely to have required sprouting by neurons rather than mere enabling of existing but formerly silent connections. The newfound cortical plasticity in adults was welcomed by cognitive neuroscientists, and had implications for recovery from brain damage and the role that sensory experience plays in it.

Equivalent experiments were soon performed in area V1, the primary visual cortex[7,8]. Initial reports showed that retinal lesions resulted, after several months, in the filling-in of the representation of the affected zone in V1 with input from the region of the retina directly surrounding the damage. It was also reported that cortical areas near the boundary of the region where the input was damaged showed changes within minutes of the lesion (for example, an increase in the size of the receptive fields there).

Over the years, however, the remapping picture in both somatosensory and visual

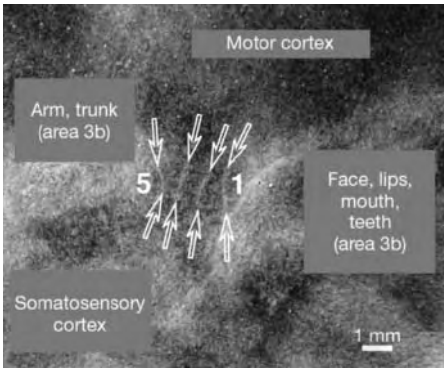


Figure 1 | Representation of fingers in the monkey somatosensory cortex. Somatosensory cortical area 3b contains a map of the entire body surface. In this myelin-stained section through the middle layers of the cortical sheet in owl monkey area 3b, thin septa (arrowed) are visible between the representation of each finger (1 is the thumb, 5 the little finger) where the density of myelin — and within-cortical connections — is reduced. These septa are not affected when peripheral nerve damage leads to reorganization of the finger maps, suggesting there are limits to cortical plasticity in the adult. (M. I. Sereno, unpublished material.)

cortex became more complicated. For example, in examining the hand representation in area 3b, Kaas and colleagues[9,10] found that the representations of the sensitive undersides of individual fingers were separated by narrow regions with reduced myelination, an indication of reduced within-cortical connections (Fig. 1). The locations of these septa, however, were not affected by the long-term loss of sensory input that resulted in somatosensory map reconfigurations in the same animals. That the septa appear at all suggests that somatosensory representations are normally quite stable despite variable experience; that they persist after peripheral damage suggests that some aspects of within-cortical connections are not plastic in the adult.

In subsequent experiments with retinal lesions in cats, stages in the input pathway were examined to try to determine the main site of the plastic changes. The cortical filling-in of V1 occurred despite a persistent dead zone in the preceding stage, the dorsolateral geniculate nucleus, and basically unaltered projections from there to V1[11]. This suggested that horizontal within-cortical connections are the main plastic element. But the filled-in cortical region contained neurons with poorer contrast sensitivity[12], and in monkeys there was a long-term reduction in cytochrome oxidase staining[13] (an indication of reduced neuronal activity). Together, these features indicated that whatever reorganization had occurred was insufficient to support normal activity, even after a long recovery.

In their work, on monkeys, Smirnakis *et al.*[1] used both functional magnetic resonance imaging (fMRI) and microelectrode recording. To their own surprise, their results came down strongly against any substantial reconfiguration of the V1 map — their fMRI data showed that the 'hole' created in V1 by a retinal lesion was still there after more than seven months; and their microelectrode data suggested that the responses inside the hole were correspondingly weak.

How can the disparity with earlier results be explained? One possibility is that the previous microelectrode recordings may have been subject to 'single-unit' selection bias — with a microelectrode, it is difficult to estimate the proportion of neurons that *do not* respond to a stimulus. Although fMRI has relatively poor spatial and temporal resolution compared with single-unit recordings, like optical recording (and the anatomical cytochrome oxidase stain) it avoids the selection bias that may have resulted in an overestimate of the neurons that responded. Another advantage of fMRI is that it samples the whole brain. Subsequent reports using this method may uncover how higher visual areas react to the hole in V1 — a topic of great interest, because visual experience depends on those areas too.

Finally, these results[1] contrast with an account[14] of cortical reorganization in a human, assessed 20 years after retinal degeneration, which showed substantial filling-in of activity in what used to be the representation of the fovea in V1. In addition to the much longer recovery time, the lesion included the entire fovea instead of being situated to one side of it, and the subject was conscious. It remains to be seen what factor or factors can explain the difference.

The possibility of plasticity in the adult cortex plays on the hope that, if one only tries hard enough, it is possible to overcome neural adversity. But hope must not obscure the data — for now, it seems that one must try for a very long time indeed if the area in question is V1. ■

Martin I. Sereno is in the Department of Cognitive Science, University of California at San Diego, La Jolla, California 92093-0515, USA.
e-mail: sereno@cogsci.ucsd.edu

1. Smirnakis, S. M. *et al. Nature* **435,** 300–307 (2005).
2. Basser, L. S. *Brain* **85,** 427–460 (1962).
3. Curtiss, S. & de Bode, S. *Brain Lang.* **86,** 193–206 (2003).
4. Merzenich, M. M. *et al. Neuroscience* **8,** 35–55 (1983).
5. Gaze, R.M. *Br. Med. Bull.* **30,** 116–121 (1974).
6. Pons, T. *et al. Science* **252,** 1857–1860 (1991).
7. Kaas, J. H. *et al. Science* **248,** 229–231 (1990).
8. Gilbert, C. D. & Wiesel, T. N. *Nature* **356,** 150–152 (1992).
9. Jain, N., Catania, K. C. & Kaas, J. H. *Cereb. Cortex* **8,** 227–236 (1998).
10. Qi, H.-X. & Kaas, J. H. *J. Comp. Neurol.* **477,** 172–187 (2004).
11. Darian-Smith, C. & Gilbert, C. D. *J. Neurosci.* **15,** 1631–1647 (1995).
12. Chino, Y. M. *et al. J. Neurosci.* **15,** 2417–2433 (1995).
13. Horton, J. C. & Hocking, D. R. *J. Neurosci.* **18,** 5433–5455 (1998).
14. Baker, C. I. *et al. J. Neurosci.* **25,** 614–618 (2005).

©2005 Nature Publishing Group

TIMEKEEPING

Light-insensitive optical clock

Thomas Udem

Interactions between trapped neutral atoms have prevented their use as the ultimate frequency standard in optical clocks. A clever trapping scheme circumvents this problem and may push timekeeping to new limits.

Almost all clocks consist of two essential parts: an oscillator with periodically occurring events, and a counter to keep track of those events. For example sundials, the oldest known clocks, use Earth's rotation as the oscillator and a human as the counter. Pendulum clocks, first made by Christiaan Huygens in 1656, have a pendulum oscillator and an automatic mechanical counter to display time. Coming right up to date, on page 321 of this issue, Takamoto and co-workers[1] demonstrate the latest innovation in timekeeping.

Modern clocks use fast oscillators and electronic counters. In the quartz clocks developed in the 1920s, a small crystal is excited to vibrate typically some 10,000 times per second. In atomic clocks, the nucleus of an atom precesses in the field created by its electrons. Caesium atomic clocks use a standard detection method to follow these precessions[2] and to feed the resulting oscillatory signal to an electronic counter, which advances the second hand of the clock precisely every 9,192,631,770 cycles. This rather odd value corresponds to the transition frequency between two energy levels of the ground state of the caesium-133 atom and has been defined as the SI unit of time since 1967. Modern caesium atomic clocks can be as accurate as one part in 10^{15}, which means that time is by far the most precisely measurable physical quantity.

Because clock instability (the variation in its 'ticking', or displayed time flow) can be reduced by increasing the oscillator frequency, it is clear how even today's most precise caesium atomic clocks may be improved: replace the caesium by an optical oscillator. The typical frequencies for such optical oscillators are some 100 or even 1,000 THz (10^{14} or 10^{15} cycles per second), and the oscillations are emitted (or absorbed) in the form of light. For a long time, this kind of oscillator was much easier to create than a correspondingly fast optical counter or clock mechanism. Efforts to build a counter that was fast enough began in the early 1960s, but it was only in 1995 that counters were developed that could reach the frequency of visible light[3]. However, these delicate and complicated counters required a number of lasers, filling several large laboratories, and so were not generally considered a serious option for a clock mechanism.

Things changed in 1999 with the introduction of an optical counter based on a pulsed femtosecond (10^{-15}-second) laser[4]. This compact device can operate continuously for days, so that counting optical oscillations became a simple task, and the first experimental optical clocks appeared[5]. A good optical oscillator requires a sharp transition that gives a strong signal with little noise and a lack of sensitivity to external perturbations. Over the years, the quest for the most suitable oscillator has been pursued along two main lines.

In the first approach, a single ion is trapped in electric fields and brought to rest by laser cooling. If both trap and ion are suitably chosen, the trapping fields will not seriously perturb the clock transition, which is triggered by a highly stable clock laser whose frequency is measured by an optical counter. Various ions are under consideration, and have produced impressive results (see for example ref. 6). But although trapped ions are largely immune to constant but unknown shifts in frequency within the system (so-called systematic uncertainties), they do not give stable enough measurements, because the signal is generated from only one ion and must be averaged over a long period. And it has not been possible to trap large numbers of ions without introducing large frequency shifts.

The second approach uses clouds of millions of neutral atoms to produce a much stronger signal, resulting in excellent stability. However, collisions between the atoms result in uncontrollable shifts in frequency. In other words, single-ion standards have high statistical but low systematic uncertainties whereas neutral-atom standards have excellent statistics but poor systematic properties.

Takamoto *et al.*[1] demonstrate a method that combines the advantages of both atomic standards. Clouds of neutral strontium atoms are held in a different kind of trap formed at the crests (anti-nodes) of a standing wave produced by a laser (Fig. 1a). In previous attempts, the laser intensity that is needed to keep neutral atoms in place significantly perturbed the transition frequency. Measuring the intensity of the trapping laser to compensate for this negative effect does not work because intensity, unlike frequency or time, cannot be measured sufficiently accurately. Instead, the authors adjust the trapping-laser frequency — which would otherwise be unimportant — so that it shifts the upper and lower energy levels of the clock transition by exactly the same amount, leaving the clock-transition frequency unaltered (Fig. 1b). The frequency of the trapping laser is then set to the correct value — to within 1 MHz, which is sufficient to suppress its effect on the clock transition to less than 0.001 Hz at a clock frequency of 429 THz.

All-optical clocks are predicted to keep time within a fractional uncertainty of 10^{-18}. Such clocks could allow fundamental physical theories to be verified at a new level. Testing general relativity and quantum electrodynamics, and detecting slow variations in the fundamental constants predicted by some theories, rely on the most precise clocks. At the 10^{-18} level of accuracy, the gravitational redshift of two clocks differing in their distance from Earth's centre by only 1 cm would become measurable. Moreover, the optical equivalent of very-long-baseline interferometry may become feasible, allowing higher-resolution astronomical images to be produced by combining data collected by widely separated telescopes. These clocks may also help to improve technologies, such as satellite navigation and broadband network synchronization, that are based on precision timing. ■

Thomas Udem is at the Max-Planck-Institut für Quantenoptik, Hans-Kopfermann-Strasse 1, 85748 Garching, Germany.
e-mail: thomas.udem@mpq.mpg.de

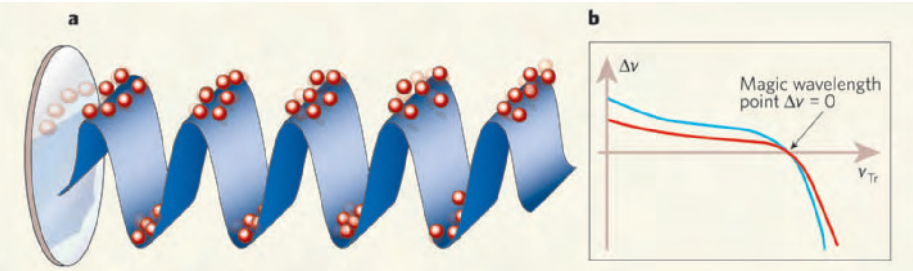


Figure 1 | Details of the optical oscillator proposed by Takamoto *et al.*[1]. **a**, Strontium atoms (red) are trapped at the crests of a standing wave formed by a laser (blue), which is reflected back onto itself by a mirror. A second laser beam (not shown) measures the clock transition. **b**, By choosing the appropriate trapping-laser frequency ν_{Tr}, the perturbation $\Delta\nu$ of the clock transition can be eliminated. The graph shows $\Delta\nu$ as a function of ν_{Tr} high (blue) and low (red) trapping-laser intensity. At the 'magic wavelength', the clock-transition perturbation vanishes, irrespective of the trapping-laser intensity, which is hard to control with the required accuracy.

1. Takamoto, M., Hong, F.-L., Higashi, R. & Katori, H. *Nature* **435,** 321–324 (2005).
2. Ramsey, N. F. *Phys. Rev.* **78,** 695–699 (1950).
3. Schnatz, H. *et al. Phys. Rev. Lett.* **76,** 18–21 (1996).
4. Udem, Th. *et al. Nature* **416,** 233–237 (2002).
5. Diddams, S. A. *et al. Science* **293,** 825–828 (2001).
6. Margolis, H. S. *et al. Science* **306,** 1355–1358 (2004).

©2005 Nature Publishing Group

NATURE|Vol 435|19 May 2005

OBITUARY

Saunders Mac Lane (1909–2005)

Outstanding mathematician and mentor who invented category theory.

Saunders Mac Lane, who died on 14 April, was one of the few mathematicians to create a major new concept that has had a lasting effect on the subject. In 1945, he and Sammy Eilenberg introduced the basic ideas of category theory, often described as a language of mathematics that allows the description of transformations from one area of the subject into another by distilling their common properties. More specifically, one can look at common properties of mathematical objects such as algebraic structures (groups, fields or rings) or topological spaces by studying structure-preserving mappings — 'morphisms' — between such objects. For example, in the case of vector spaces as mathematical objects, the structure-preserving mappings are the linear transformations, and one can study the 'category' of vector spaces using linear mappings as the morphisms between objects of the category. Concentrating on common properties of these structures by stripping away the non-essential aspects of a problem clears the path to new results.

Mac Lane was born in Norwich, Connecticut, and studied at Yale University and the University of Chicago, where he obtained a master's degree in 1931. After doctoral studies at the University of Göttingen in Germany, and appointments first as an assistant and then as a full professor at Harvard University, he returned to the University of Chicago in 1947. He remained there for the rest of his career, being appointed emeritus professor in 1982.

Perhaps responding to some early comments on category theory, Mac Lane wrote in his autobiography: "At the time, we sometimes called our subject 'general abstract nonsense'. We didn't really mean the nonsense part, and we were proud of its generality." Later, Mac Lane wrote *Categories for the Working Mathematician*, the title of which reflects his ability to convey humorously a serious intention — in this case, to make an abstract idea relevant.

Mac Lane's innovative contributions to mathematics had two roots. One was his interest in what he called 'universal knowledge', which he recognized as a student at Yale. With a wry awareness of the problems this posed, he later wrote: "I believed that a proper academic ought to know everything, but of course, I never did succeed in mastering universality, nor do I know exactly what this might have meant." The other root was his firm belief that mathematics has to be developed from the specific to the general and that abstraction is a tool, not a goal. This belief is reflected in the books that he wrote. *A Survey of Modern Algebra*, written 50 years ago with his co-instructor at Harvard, Garrett Birkhoff, is still used as a text; and his classic monograph, *Homology*, in which every general concept is presented starting from a specific mathematical situation, still serves as an introduction and reference to that subject.

Mac Lane's impact on the mathematical community went beyond his contributions as a researcher. According to the online 'Mathematics Genealogy Project', he had 39 students and 1,028 'descendants'. He recognized talent and let it flourish with modest interference. Among his protégés were Irving Kaplansky, his first PhD student at Harvard, who later headed the Mathematical Sciences Research Institute at the University of California, Berkeley; and John Thompson, winner in 1970 of the highest mathematical honour, the Fields medal, for his work on the so-called 'odd-order' problem, which solved the Burnside conjecture that every finite simple group of non-prime order must have even order.

Recalling his work with Thompson, Mac Lane wrote: "By the end of the second quarter, in which I had tended to emphasize infinite groups, I had essentially come to the end of my knowledge on group theory. Thompson, who was in the course, came to me to say that he wished to write a thesis on group theory. I encouraged him, but did not trouble to say that my own knowledge of the subject was somewhat limited. But not to worry — I arranged for eminent theorists, such as Richard Brauer, Reinhold Baer, and Marshall Hall, to visit Chicago. Each Saturday morning, I listened to Thompson tell me what he had been up to with groups; the subject fitted his interest, and he chose his own problems."

This spirit of guidance and selfless encouragement probably grew out of Mac Lane's time in Göttingen, then the Mecca of mathematics, in the early 1930s, where he became immersed in an atmosphere of intellectual challenge and excitement that formed his views. It was there that he received his PhD with a thesis on "Abbreviated proofs in logic calculus". Following ideas from Alfred North Whitehead and Bertrand Russell, he worked on formalizing mathematical proof. The subject did not impress his advisers Hermann Weyl and Paul Bernays, and in Mac Lane's own words: "My thesis did not attract any following, nor did it have any influence on subsequent studies of mechanized proof."

UNIV. CHICAGO

In his later career, Mac Lane took his responsibilities to the wider scientific community very seriously, always devoting his full energies to the task at hand. He was elected to the National Academy of Sciences in 1949, and served as vice-president of the academy and of the American Philosophical Society, and as president of the American Mathematical Society. During his presidency of the Mathematical Association of America in the 1950s, he began the association's first efforts to improve the teaching of modern mathematics. He was also a member of the National Science Board (1974–80), which provided science policy advice to the US government. In 1976, he led a delegation of mathematicians to the People's Republic of China to examine the conditions affecting the development of mathematics there. In 1989, he received the highest US award for scientific achievement, the National Medal of Science.

During the past few years, Mac Lane was engaged in writing his autobiography, and I have been privileged to be a part of that process. In the book, he wanted to convey to a new generation of mathematicians the joy of the creative process and the excitement that he experienced in being part of history. ■

Klaus Peters

Klaus Peters is at A K Peters Ltd, 888 Worcester Street, Suite 230, Wellesley, Massachusetts 02482-3717, USA. He is the publisher of *Saunders Mac Lane — A Mathematical Autobiography*, which will appear at the end of May.

e-mail: klaus@akpeters.com

©2005 Nature Publishing Group

BRIEF COMMUNICATIONS

Red enhances human performance in contests

Signals biologically attributed to red coloration in males may operate in the arena of combat sports.

Red coloration is a sexually selected, testosterone-dependent signal of male quality in a variety of animals[1–5], and in some non-human species a male's dominance can be experimentally increased by attaching artificial red stimuli[6]. Here we show that a similar effect can influence the outcome of physical contests in humans — across a range of sports, we find that wearing red is consistently associated with a higher probability of winning. These results indicate not only that sexual selection may have influenced the evolution of human response to colours, but also that the colour of sportswear needs to be taken into account to ensure a level playing field in sport.

Although other colours are also present in animal displays, it is specifically the presence and intensity of red coloration that correlates with male dominance and testosterone levels[3–5]. In humans, anger is associated with a reddening of the skin due to increased blood flow[7,8], whereas fear is associated with increased pallor in similarly threatening situations[9]. Hence, increased redness during aggressive interactions may reflect relative dominance. Because artificial stimuli can exploit innate responses to natural stimuli[6,10], we tested whether wearing red might influence the outcome of physical contests in humans.

In the 2004 Olympic Games, contestants in four combat sports (boxing, tae kwon do, Greco–Roman wrestling and freestyle wrestling) were randomly assigned red or blue outfits (or body protectors). If colour has no effect on the outcome of contests, the number of winners wearing red should be statistically indistinguishable from the number of winners wearing blue. However, we found that for all four competitions, there is a consistent and statistically significant pattern in which contestants wearing red win more fights ($\chi^2 = 4.19$, d.f. = 1, $P = 0.041$; Fig. 1a). This result is remarkably consistent across rounds in each competition, with 16 of 21 rounds having more red than blue winners, and only four rounds having more blue winners (sign test, $P = 0.012$). The effect is the same across the weight classes in each sport: 19 of 29 classes had more red winners, with only six classes having more blue winners (sign test, $P = 0.015$). (For methods, see supplementary information.)

Given the undoubted role of other factors, such as skill and strength, it is likely that the red advantage will determine the outcome only in relatively symmetric contests. That is, wearing red presumably tips the balance between losing and winning only when other factors are fairly equal. We found that this is indeed the case: only in contests between individuals of similar ability were there significantly more red than blue winners ($\chi^2 = 6.07$, d.f. = 1, $P = 0.014$), with the red advantage seeming to decline as asymmetries in competitive ability increase (Fig. 1b). Hence, although the effect is significant for the pooled data shown in Fig. 1a, this is due principally to the results for relatively symmetric contests.

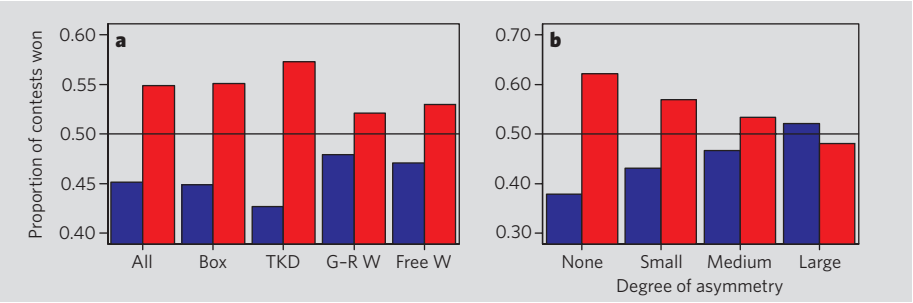


Figure 1 | Influence of colour of sporting attire on the outcome of competitive sports. a, Proportion of contests in Olympic combat sports won by competitors wearing red (right bars) or blue (left bars) outfits for all sports combined and for the individual sports of boxing (Box), tae kwon do (TKD), Greco-Roman wrestling (G–R W) and freestyle wrestling (Free W). **b**, Proportion of contests won by competitors wearing red or blue given different degrees of relative ability (asymmetry) in the two competitors in each bout. No significant differences exist between the number of red and blue wins for contests with small ($\chi^2 = 2.21$, d.f. = 1, $P = 0.14$), medium ($\chi^2 = 0.47$, d.f. = 1, $P = 0.50$) or large asymmetries ($\chi^2 = 0.21$, d.f. = 1, $P = 0.64$) in competitive ability. Black lines at 0.5 indicate the expected proportion of wins by red or blue under the null hypothesis that colour has no effect on contest outcomes. For details of data collection and analyses, see supplementary information.

These results indicate that artificial colours may influence the outcome of physical contests in humans. A preliminary analysis of the results of the Euro 2004 international soccer tournament, in which teams wore shirts of different colours in different matches, suggests that wearing red may also bestow an advantage in team sports and when opponents wear colours other than blue. We compared the performance of five teams that wore a predominantly red shirt against their performance when wearing a different shirt colour (four played their other matches in white, one in blue). We found that all five had better results when playing in red (paired t-test, $t = -3.15$, d.f. = 4, $P = 0.034$), largely as a result of scoring more goals ($t = -2.98$, d.f. = 4, $P = 0.041$) (further details are available from the authors). Hence, colour of sportswear may affect outcomes in a wide variety of sporting contexts.

Colour is thought to influence human mood, emotions and expressed aggression, and is a recognized element of signalling in competitive interactions in many non-human species. But it has not hitherto been suspected to be a factor in human contests. Given the ubiquity of aggressive competition throughout human societies and history, our results suggest that the evolutionary psychology of colour is likely to be a fertile field for further investigation. The implications for regulations governing sporting attire may also be important.

Russell A. Hill, Robert A. Barton
Evolutionary Anthropology Research Group, University of Durham, Durham DH1 3HN, UK
e-mail: r.a.hill@durham.ac.uk

1. Milinski, M. & Bakker, T. C. M. *Nature* **344,** 330–333 (1990).
2. Waitt, C. *et al. Proc. R. Soc. Lond. B* **270,** S144–S146 (2003).
3. Setchell, J. M. & Wickings, E. J. *Ethology* **111,** 25–50 (2005).
4. Pryke, S. R., Andersson, S., Lawes, M. J. & Piper, S. E. *Behav. Ecol.* **13,** 622–631 (2002).
5. Andersson, S., Pryke, S. R., Ornborg, J., Lawes, M. J. & Andersson, M. *Am. Nat.* **160,** 683–691 (2002).
6. Cuthill, I. C. *et al. Proc. R. Soc. Lond. B* **264,** 1093–1099 (1997).
7. Drummond, P. D. & Quah, S. H. *Psychophysiology* **38,** 190–196 (2001).
8. Darwin, C. *The Expression of the Emotions in Man and Animals* (Murray, London, 1872).
9. Drummond, P. D. *Pers. Indiv. Diff.* **23,** 575–582 (1997).
10. Ryan, M. J. in *Behavioural Ecology* (eds Krebs, J. R. & Davies, N. B.) 179–202 (Blackwell Science, Oxford, 1997).

doi: 10.1038/435293a
Supplementary information accompanies this communication on *Nature*'s website.
Competing financial interests: declared none.

©2005 Nature Publishing Group

ARTICLES

Three-dimensional deformation caused by the Bam, Iran, earthquake and the origin of shallow slip deficit

Yuri Fialko[1], David Sandwell[1], Mark Simons[2] & Paul Rosen[3]

Our understanding of the earthquake process requires detailed insights into how the tectonic stresses are accumulated and released on seismogenic faults. We derive the full vector displacement field due to the Bam, Iran, earthquake of moment magnitude 6.5 using radar data from the Envisat satellite of the European Space Agency. Analysis of surface deformation indicates that most of the seismic moment release along the 20-km-long strike-slip rupture occurred at a shallow depth of 4–5 km, yet the rupture did not break the surface. The Bam event may therefore represent an end-member case of the 'shallow slip deficit' model, which postulates that coseismic slip in the uppermost crust is systematically less than that at seismogenic depths (4–10 km). The InSAR-derived surface displacement data from the Bam and other large shallow earthquakes suggest that the uppermost section of the seismogenic crust around young and developing faults may undergo a distributed failure in the interseismic period, thereby accumulating little elastic strain.

Over the past decade, new information about the near-field deformation due to several large shallow earthquakes was obtained with the help of the space-borne interferometric Synthetic Aperture Radar (InSAR) measurements[1–3]. Interpretations of the spatially continuous SAR data from the best-documented seismic events including the $M_w = 7.3$ Landers[4], the $M_w = 7.6$ Izmit[5,6], and the $M_w = 7.1$ Hector Mine[3,7,8] earthquakes all reveal the maximum seismic moment release in the middle of the seismogenic layer (at average depths of 4–6 km). Whereas a gradual decay in the coseismic slip at the bottom of the seismogenic layer is probably compensated by postseismic and interseismic strain accumulation, and is reasonably well understood[9–11], the apparent discrepancy between slip in the middle and shallow parts of the seismogenic layer remains enigmatic.

The uppermost few kilometres of the brittle crust are known to have mechanical properties that differ from those of the rest of the upper crust. In particular, the shallow layer has a higher density of cracks, pores and voids[12], a higher coefficient of friction[13], and may exhibit velocity-strengthening behaviour[14]. The latter may explain why the coseismic slip may be impeded in the shallow crust, but it is not clear how the resulting deficit of shallow slip is accommodated throughout the earthquake cycle. Steady-state shallow creep has been inferred from the SAR data in some localities (such as on the southern section of the San Andreas fault[15]), but more often has not been observed. The remaining alternatives are episodic shallow creep, shallow postseismic afterslip, or a distributed inelastic failure of the shallow crust, either during earthquakes[3], or in the interseismic period. The mechanisms of accumulation and release of stress and strain in the shallow seismogenic crust are of interest because most of the seismic and geodetic measurements of deformation are done at the surface or in shallow boreholes. The mode of deformation and the state of stress in the uppermost crust are also important for predictions of the intensity of ground shaking, and the associated seismic hazards in the vicinity of large seismogenic faults.

Here we report on deformation associated with the $M_w = 6.5$ Bam earthquake in Iran determined using the SAR data from the ERS and Envisat satellites of the European Space Agency. The Bam earthquake is the first large ($M_w > 6$) shallow earthquake for which the decorrelation of the radar images does not prevent measurements of surface displacements across the earthquake rupture, thereby allowing robust insights into the problem of the shallow slip deficit.

The $M_w = 6.5$ Bam earthquake occurred on 26 December 2003, in southeastern Iran within a diffuse boundary between the Arabian and Eurasian plates. It was one of the deadliest earthquakes in the region's history, with an estimated several tens of thousands of casualties. The earthquake rupture occurred directly below the town of Bam (Supplementary Fig. S1), causing nearly complete destruction of old un-reinforced (predominantly mudbrick) and modern buildings. Teleseismic (from US Geological Survey, USGS, and the Harvard Centroid Moment Tensor, CMT, project) and

Table 1 | The coseismic interferometric pairs used in this study

Pair number	Acquisition dates	Orbit	$B_\perp$ (m)	Sensor
Coseismic				
IP1	2003/12/03 to 2004/02/11	Descending	2	Envisat
IP2	2003/11/16 to 2004/01/25	Ascending	30	Envisat
Postseismic				
IP3	2004/01/25 to 2004/02/29	Ascending	34	Envisat
Preseismic				
IP4	1992/12/06 to 1996/04/02	Descending	118	ERS-1
IP5	1992/07/19 to 1996/04/03	Descending	83	ERS-1,2
IP6	1993/09/12 to 1996/05/08	Descending	44	ERS-1,2
IP7	1992/11/01 to 1996/05/07	Descending	26	ERS-1
IP8	1992/07/19 to 1996/04/02	Descending	38	ERS-1
IP9	1992/07/19 to 1997/05/28	Descending	7	ERS-1,2
IP10	1993/09/12 to 1998/09/30	Descending	9	ERS-1,2
IP11	1996/04/02 to 1999/03/24	Descending	14	ERS-1,2
IP12	1996/05/07 to 1999/06/02	Descending	11	ERS-1,2

$B_\perp$ is the across-track separation between the repeated satellite orbits.

[1]Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California 92093, USA. [2]Seismological Laboratory, Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA. [3]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA.

© 2005 Nature Publishing Group

preliminary aftershock data[16] indicate a strike-slip mechanism with a right-lateral slip on a nearly vertical fault. The epicentral area of the earthquake was imaged by the ASAR (advanced SAR) instrument on board the Envisat satellite within several weeks of the seismic event, with acquisitions available from the ascending and descending satellite orbits. Table 1 lists the radar acquisitions used in this study.

We generated four independent projections of the coseismic displacement field using differences in the radar phase[17] and the radar amplitude[8,18] before and after the earthquake from both the ascending and descending orbits (see Methods). The radar line of sight (LOS) displacements and the pixel offsets along the satellite tracks are shown in Fig. 1a, b and Fig. 1c, d, respectively. The strike-slip mechanism and the north–south orientation of the Bam rupture are optimal in that they maximize the azimuthal pixel offsets (AZO). The correlation of the radar images is exceptionally good, presumably owing to arid conditions and sparse vegetation. The only decorrelated areas, around the northern end of the rupture, reflect the massive destruction (and possibly postearthquake rescue and remedy activities) in the town of Bam.

Three-dimensional coseismic offsets due to the Bam earthquake

We combine the four projections of surface deformation (Fig. 1a–d) to deduce the full three-dimensional vector displacement caused by the Bam earthquake[4,8]. Figures 1e and f show the vertical and horizontal components of the coseismic deformation, respectively. The data pairs from the ascending and descending orbits include several weeks of possible postseismic relaxation. Postseismic deformation is probably negligible compared to the coseismic offsets, as observations of postseismic deformation due to large strike-slip earthquakes elsewhere show[19–21]. Therefore it is reasonable to believe that the data shown in Fig. 1 are dominated by the coseismic deformation.

The location of the earthquake rupture is readily identifiable in the horizontal displacement map as the north–south striking plane of symmetry between the butterfly-shaped lobes of the coseismic offsets (Fig. 1f). Such a spatial pattern, as well as the antisymmetry of both the horizontal and vertical displacements with respect to the fault plane, is predicted by elastic dislocation models of the earthquake source[4,8,22]. The coseismic displacement field inferred from the Envisat ASAR data reveals somewhat greater displacement amplitudes on the eastern side of the fault, implying either a contrast in the elastic moduli between the eastern and western sides of the fault, or a small eastward deviation of the fault plane from the vertical. To determine the subsurface fault structure we inverted the interferometric and the azimuthal offset data (Fig. 1a–d) for the fault geometry and slip distribution (see Methods).

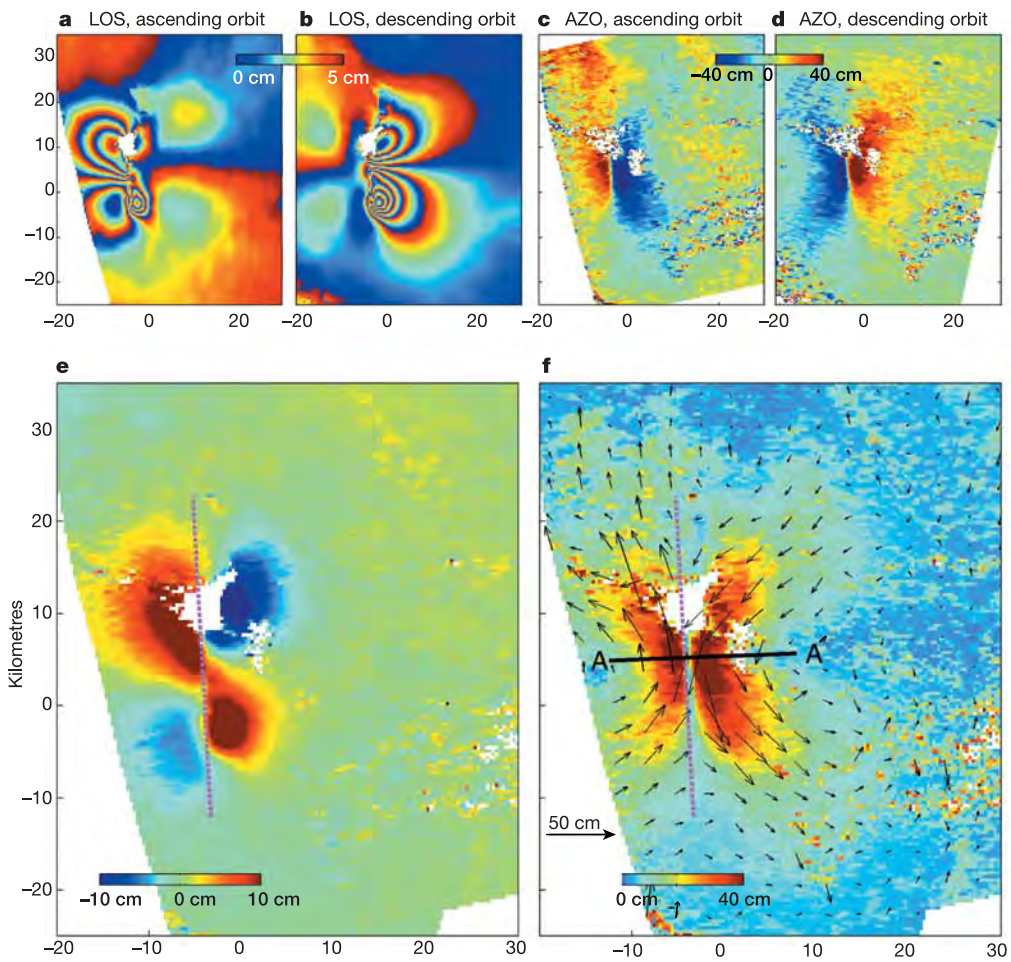


Figure 1 | Coseismic deformation caused by the Bam earthquake as imaged by the Envisat ASAR data. The coordinate axes are in kilometres, with the origin at 58.4° E, 29° N. Colours denote displacements in centimetres. **a**, Interferogram for the time period 16 November 2003 to 25 January 2004, ascending orbit. **b**, Interferogram for the time period 3 December 2003 to 11 February 2004, descending orbit. **c**, Azimuthal offsets, ascending orbit. **d**, Azimuthal offsets, descending orbit. **e**, **f**, Vertical (**e**) and horizontal (**f**) components of the surface displacement field derived from the ASAR data (**a**–**d**). Arrows show the subsampled horizontal displacements. Dashed line shows the surface projection of the fault plane inferred from the inverse modelling of the ASAR data.

© 2005 Nature Publishing Group

Supplementary Fig. S2 shows the slip model that best explains the ASAR data, and Supplementary Fig. S3 shows the model predictions and the data residuals. The best-fitting model indicates predominantly right-lateral displacements having a maximum amplitude of about 2 m at a depth of 3 to 7 km. The geodetic moment determined by summation of the dislocation potencies (area times slip), and multiplying the sum by the typical value of the shear modulus of the Earth's crust (30 GPa) is of the order of $(6–8) \times 10^{18}$ N m. This corresponds to the moment magnitude of 6.5–6.6, in excellent agreement with the seismically determined values[16]. A major peculiarity of the inferred slip model is that the maximum moment release occurred at a fairly small depth (~4 km), yet the slip did not reached the surface (Supplementary Fig. S2). The lack of surface rupture due to the Bam earthquake is clear from the continuity of fringes in the radar interferograms (Fig. 1a, b), and is confirmed by field investigations[16,23].

Figure 2 shows the absolute value of horizontal displacements from a 4-km-wide swath across the central part of the Bam rupture (see Fig. 1f). We can see from Fig. 2 that the surface offsets on the surface trace of the Bam rupture do not exceed a few centimetres, and are much smaller than the maximum horizontal displacements of 0.5–0.6 m that occur at a distance of about 1.5 km away from the rupture trace. Small-to-moderate ($M_w < 6$) crustal earthquakes typically do not break the surface because they nucleate at depth, and have a characteristic rupture size that is small compared to the thickness of the brittle layer[24,25]. This is not the case for the Bam rupture, which has a characteristic horizontal dimension of about 20 km (Supplementary Figs S1 and S2), that is, sufficient to saturate the entire upper crust.

Nature of the shallow slip deficit

Given that the crustal strength decreases toward the surface[13,26], the termination of slip in the uppermost crust indicates either significant velocity strengthening or negligible preseismic elastic strain at depths of more than 2–3 km (or a combination of the above). Qualitatively, the distribution of slip due to the Bam earthquake is similar to the distributions inferred for other large strike-slip earthquakes for which high-quality geodetic data are available. Figure 3 shows the average seismic potency per unit length of rupture for the Bam (this study), Landers[4], Izmit[6] and Hector Mine[3] earthquakes. In all cases, the maximum release of seismic moment occurs in the middle of the brittle layer, and decreases toward the surface. (The near-surface decrease in the coseismic slip may be less apparent for the Izmit earthquake because of the low resolution of both the available slip models and the data, owing to significant decorrelation of the ground around the earthquake rupture. Other finite-source models of the Izmit earthquake show a more pronounced slip deficit[5,27].)

Assuming that the earthquake rupture is an ergodic process (that is, global spatial sampling is equivalent to local temporal sampling over many earthquake cycles), the results shown in Fig. 3 pose a dilemma: either the elastic dislocation models are inadequate for interpretation of the coseismic deformation data, or much of the stress release in the shallow crust occurs aseismically. The first hypothesis implies that the observed inflection (that is, the change in sign of the second spatial derivative), or even non-monotonic behaviour (that is, the change in sign of the first spatial derivative, as is the case for the Bam earthquake) of the surface displacements in the near field of the seismic rupture is due to an essentially inelastic response of the uppermost few kilometres of the brittle crust[3]. In this case, the shallow slip deficit is an artefact of inverse models that are based on elastic solutions[28,29], and the surface slip need not be systematically less than the maximum slip at depth. Unfortunately, the surface displacements inferred from previous SAR studies cannot be directly compared to the fault offsets measured in the field because of the decorrelation of the radar images around the earthquake rupture[3,4,6,8]. The data from the Bam earthquake are unambiguous in that the displacements can be continuously traced across the fault, indicating no slip in the shallow crust (Figs 1 and 2, and Supplementary Fig. S2). The data shown in Figs 1f and 2 indicate that the assumption of no coseismic slip deficit implies that the inelastic deformation in the shallow crust is distributed within a ~3-km-wide shear zone. This implication is not supported by inspections of the radar phase coherence in the earthquake epicentral area, which show a rather localized zone of surface damage of the order of tens to hundreds of metres wide, and field observations of microcracking and small-scale offsets limited to the rupture trace of the fault[23].

Regardless of whether the coseismic slip in the top few kilometres of the seismogenic layer is inhibited by the velocity-strengthening behaviour, or low shear stress, the observed slip deficit apparently has to be accommodated aseismically, as intermediate-size earthquakes as well as intense microseismicity in the top 2–3 km of the crust are extremely rare. It has been proposed that the coseismically induced

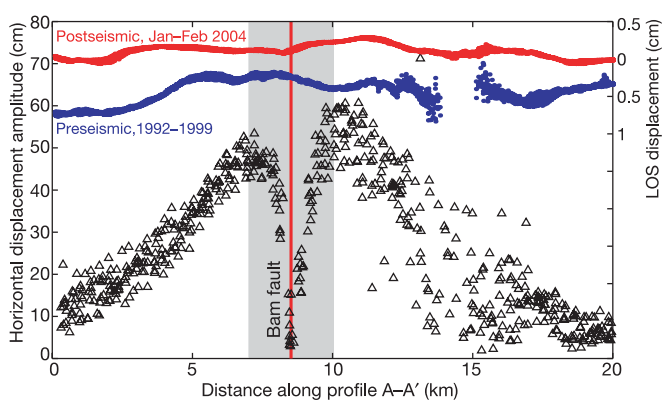


Figure 2 | Displacements along a profile A–A′ perpendicular to the Bam earthquake rupture (Fig. 1). Black triangles denote the absolute value of coseismic displacements (left axis), red dots denote the postseismic LOS displacements that occurred over a time period of one to two months after the earthquake, and blue dots denote the preseismic LOS displacements that occurred over a time period between 1992 and 1999 (right axis). The red vertical line denotes the position of the rupture trace deduced from the phase correlation map. The grey bar marks a 3-km-wide zone between the maxima in the amplitude of horizontal displacements on both sides of the fault.

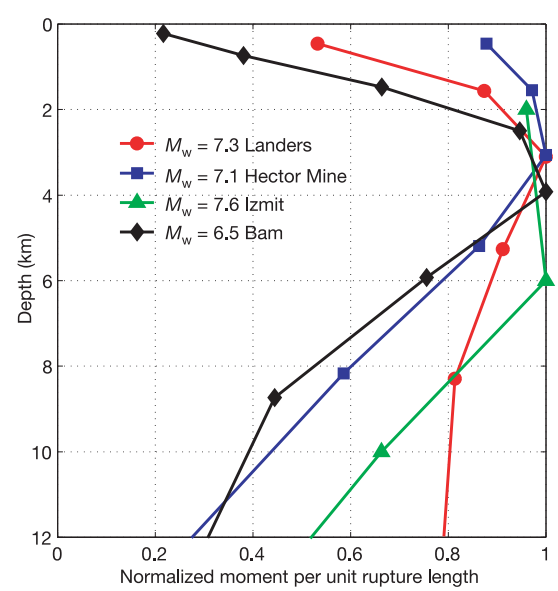


Figure 3 | Distribution of seismic potency $\bar{P}$ averaged along the fault length L. $\bar{P}(z) = \int_0^L P(x,z)\mathrm{d}x / \max\left(\int_0^L P(x,z)\mathrm{d}x\right)$, as a function of depth z, is shown for several large strike-slip earthquakes.

© 2005 Nature Publishing Group

stress changes may give rise to an accelerated stable slip in the shallow velocity-strengthening layer[30]. Although a significant shallow afterslip has been documented on faults that are prone to creep in the interseismic period[31,32], more often it has not been observed[6,19–21]. Our analysis of InSAR data over the time period of two months following the Bam earthquake also does not reveal any shallow afterslip on the earthquake rupture (Fig. 2). Preliminary InSAR results spanning a time period of ten months after the earthquake confirm that the slip deficit was not relieved in the postseismic period.

Previous studies have shown that some shallow slip may be triggered on faults as a result of nearby earthquakes[8,33,34]. However, the amount of such triggered slip (integrated over the earthquake cycle) is unlikely to account for the estimated slip deficit of the order of metres. Another possibility is that the localized shallow slip occurs at a nearly constant rate during the interseismic period. Some faults are inferred to undergo a quasi-steady-state shallow creep[15,26,35]. However, decade-long InSAR observations in the Eastern California Shear Zone[21,36] and other seismically active areas around the world suggest that the steady shallow creep is an exception rather than a rule, especially for immature and infrequently slipping faults such as the Bam rupture. Analysis of the ERS SAR data spanning the time period between 1992 and 1999 lends further support to this conclusion (see Fig. 2 and Table 1).

We propose that the shallow slip deficit results from a distributed inelastic deformation within the uppermost few kilometres of the Earth's crust, occurring predominantly during the interseismic period. The non-brittle long-term behaviour of the uppermost crust is well known from field studies of compressional tectonics (in particular, blind thrust faults)[37,38]. For strike-slip faults, the interseismic deformation may involve a predominantly elastic deformation of the upper crust below ~2–3 km, and predominantly inelastic deformation of the uppermost layer owing to folding, granular flow, or some other distributed failure mechanism. Alternatively, the uppermost crust may be partially decoupled from the seismogenic layer, for example, by a low friction interface. In both cases the infrequently slipping strike-slip faults that rarely break the surface may be very difficult to detect from geologic and palaeoseismologic observations[23]. The non-localized nature of near-surface deformation is consistent with velocity-strengthening friction and low absolute strength of the poorly consolidated uppermost crust[26], and may explain the 'flower structures' associated with major strike-slip faults[39,40].

According to our hypothesis the shallow crust can be either weak or strong (for example, able to support stresses predicted by Byerlee's law), but may not accumulate significant elastic strain owing to the slow tectonic loading. At the same time, it might deform elastically on short timescales (corresponding to the coseismic deformation, for example), as shown by the coseismic response of large compliant fault zones[4,41]. Whether or not the earthquake rupture reaches the surface may be controlled by the amount of the earthquake stress drop in the velocity-weakening part of the crust, and the level of the preseismic stress in the shallow layer. If the shallow layer is weak, the upward rupture propagation from the seismogenic part of the crust may give rise to a dynamic overshoot in the shallow layer.

The ongoing drilling experiment on the San Andreas fault[42] will presumably penetrate the transition between the velocity-strengthening and velocity-weakening layers within the seismogenic crust, and provide direct observational constraints on the level of stresses at which the upper sections of major strike-slip faults operate. Note that the data presented in this paper characterize deformation due to relatively young or infrequently slipping faults with small cumulative offsets. It remains to be seen whether the shallow slip deficit is typical of mature faults capable of great ($M_w > 8$) earthquakes. A good agreement between the geologic and present-day geodetic slip rates on the central section of the San Andreas fault[43] may be indicative of high localization of strain throughout the earthquake cycle. In contrast, geologically inferred slip rates on relatively young and developing faults are often systematically less than the geodetic estimates[44], consistent with our interpretation.

A distributed failure of the near-surface layer due to the secular tectonic loading implies that estimates of the depth of the brittle–ductile transition from geodetic measurements of the interseismic strain accumulation on major crustal faults may be systematically underestimated by an amount equivalent to the thickness of the anelastic surface layer. The particular modes of deformation of the shallow crust have important implications for the seismic energy release and the intensity of ground shaking in epicentral areas of moderate-to-large crustal earthquakes. Because the uppermost crust may store little potential energy of elastic deformation, it is not likely to participate in the elastic rebound, which might compound the effects of the velocity strengthening in dampening of the seismic energy radiation. This implies smaller velocities and accelerations at the Earth's surface (compared to the ideal elastic–brittle behaviour of the entire upper crust), and, consequently, reduced potential damage due to shallow earthquakes.

METHODS

Data processing and analysis. The raw SAR data were processed using the JPL/Caltech software ROI_PAC[45], and precise satellite orbits from Delft University (Netherlands)[46]. Effects of topography were removed from the interferograms using a digital elevation model produced by the Space Shuttle Radar Topography Mission[47].

The choice of the data pairs was stipulated, in particular, by (1) optimal baselines and (2) temporal proximity of the post-earthquake acquisitions from the ascending and descending orbits (see Table 1). The small interferometric baselines of the data pairs used in this study result in much better correlation of the radar images compared to the shorter time span, but larger baseline pairs[23].

A joint inversion of the interferometric and the azimuthal offset data (Figs 1a–d) was used to infer the earthquake fault location and slip distribution. For a given fault geometry, the slip distribution was found using the least-squares minimization with the non-negativity constraint on the strike-slip component of the displacement vector; no sign constraints were imposed on the dip-slip component. The optimal smoothness was determined by investigating a trade-off between the model misfit and the degree of smoothing. The fault geometry was found using multiple slip inversions and a grid search through the model parameters defining the fault location and orientation[3,4]. The best-fitting fault geometry indicates that the earthquake rupture is steeply (5° off vertical) dipping to the East (Supplementary Fig. S2), which probably explains the inferred asymmetry in the surface displacement field (Figs 1 and 2).

Received 1 October 2004; accepted 28 January 2005.

1. Massonnet, D. *et al.* The displacement field of the Landers earthquake mapped by radar interferometry. *Nature* **364,** 138–142 (1993).
2. Peltzer, G., Crampe, F. & King, G. Evidence of nonlinear elasticity of the crust from the M_w 7.6 Manyi (Tibet) earthquake. *Science* **286,** 272–276 (1998).
3. Simons, M., Fialko, Y. & Rivera, L. Coseismic deformation from the 1999 M_w 7.1 Hector Mine, California, earthquake, as inferred from InSAR and GPS observations. *Bull. Seismol. Soc. Am.* **92,** 1390–1402 (2002).
4. Fialko, Y. Probing the mechanical properties of seismically active crust with space geodesy: Study of the co-seismic deformation due to the 1992 M_w 7.3 Landers (southern California) earthquake. *J. Geophys. Res.* **109,** doi:10.1029/2003JB002756 (2004).
5. Feigl, K. L. *et al.* Estimating slip distribution for the Izmit mainshock from coseismic GPS, ERS-1, RADARSAT and SPOT measurements. *Bull. Seismol. Soc. Am.* **92,** 138–160 (2002).
6. Cakir, Z. *et al.* Coseismic and early post-seismic slip associated with the 1999 Izmit earthquake (Turkey), from SAR interferometry and tectonic field observations. *Geophys. J. Int.* **155,** 93–110 (2003).
7. Jonsson, S., Zebker, H., Segall, P. & Amelung, F. Fault slip distribution of the 1999 M_w 7.1 Hector Mine, California, earthquake, estimated from satellite radar and GPS measurements. *Bull. Seismol. Soc. Am.* **92,** 1377–1389 (2002).
8. Fialko, Y., Simons, M. & Agnew, D. The complete (3-D) surface displacement field in the epicentral area of the 1999 M_w 7.1 Hector Mine earthquake, southern California, from space geodetic observations. *Geophys. Res. Lett.* **28,** 3063–3066 (2001).
9. Tse, S. T. & Rice, J. R. Crustal earthquake instability in relation to the depth variation of frictional slip properties. *J. Geophys. Res.* **91,** 9452–9472 (1986).
10. Thatcher, W. Nonlinear strain build-up and the earthquake cycle on the San Andreas fault. *J. Geophys. Res.* **88,** 5893–5902 (1983).
11. Savage, J. & Svarc, J. Postseismic deformation associated with the 1992

 © 2005 Nature Publishing Group

M_w = 7.3 Landers earthquake, southern California. *J. Geophys. Res.* **102,** 7565–7577 (1997).
12. Manning, C. & Ingebritsen, S. Permeability of the continental crust: Implications of geothermal data and metamorphic systems. *Rev. Geophys.* **37,** 127–150 (1999).
13. Byerlee, J. Friction of rock. *Pure Appl. Geophys.* **116,** 615–626 (1978).
14. Marone, C. Laboratory-derived friction laws and their application to seismic faulting. *Annu. Rev. Earth Planet. Sci.* **26,** 643–696 (1998).
15. Lyons, S. & Sandwell, D. Fault creep along the southern San Andreas from interferometric synthetic aperture radar, permanent scatterers, and stacking. *J. Geophys. Res.* **108,** doi:10:1029/2002JB001831 (2003).
16. Tatar, M. *et al.* Aftershocks study of the 26 December 2003 Bam earthquake. *J. Seismol. Earthquake Eng.* (in the press).
17. Rosen, P. *et al.* Synthetic aperture radar interferometry. *Proc. IEEE* **88,** 333–382 (2000).
18. Michel, R. & Avouac, J.-P. Measuring ground displacements from SAR amplitude images: Application to the Landers earthquake. *Geophys. Res. Lett.* **26,** 875–878 (1999).
19. Peltzer, G., Rosen, P., Rogez, F. & Hudnut, K. Poroelastic rebound along the Landers 1992 earthquake surface rupture. *J. Geophys. Res.* **103,** 30131–30145 (1998).
20. Jacobs, A., Sandwell, D., Fialko, Y. & Sichoix, L. The 1999 (M_w 7.1) Hector Mine, California, earthquake: Near-field postseismic deformation from ERS interferometry. *Bull. Seismol. Soc. Am.* **92,** 1433–1442 (2002).
21. Fialko, Y. Evidence of fluid-filled upper crust from observations of post-seismic deformation due to the 1992 M_w 7.3 Landers earthquatek. *J. Geophys. Res.* **109,** doi:10.1029/2004JB002985 (2004).
22. Chinnery, M. A. The deformation of the ground around surface faults. *Bull. Seismol. Soc. Am.* **51,** 355–372 (1961).
23. Talebian, M. *et al.* The 2003 Bam (Iran) earthquake: Rupture of a blind strike-slip fault. *Geophys. Res. Lett.* **31,** L11611, doi:10.1029/2004GL020058 (2004).
24. Heaton, T. Evidence for and implications of self-healing pulses of slip in earthquake rupture. *Phys. Earth Planet. Inter.* **64,** 1–20 (1990).
25. Boatwright, J. Spectral theory for circular seismic sources—simple estimates of source dimension, dynamic stress drop, and radiated seismic energy. *Bull. Seismol. Soc. Am.* **70,** 1–27 (1980).
26. Savage, J. & Lisowski, M. Inferred depth of creep on the Hayward fault, central California. *J. Geophys. Res.* **98,** 787–793 (1993).
27. Delouis, B., Giardini, D., Lundgren, P. & Salichon, J. Joint inversion of InSAR, GPS, teleseismic, and strong-motion data for the spatial and temporal distribution of earthquake slip: Application to the 1999 Izmit mainshock. *Bull. Seismol. Soc. Am.* **92,** 278–299 (2002).
28. Okada, Y. Surface deformation due to shear and tensile faults in a half-space. *Bull. Seismol. Soc. Am.* **75,** 1135–1154 (1985).
29. Wang, R., Martin, F. & Roth, F. Computation of deformation induced by earthquakes in a multi-layered elastic crust—FORTRAN programs EDGRN/EDCMP. *Comp. Geosci.* **29,** 195–207 (2003).
30. Marone, C., Scholz, C. & Bilham, R. On the mechanics of earthquake afterslip. *J. Geophys. Res.* **96,** 8441–8452 (1991).
31. Williams, P., McGill, S., Sieh, K., Allen, C. & Louie, J. Triggered slip along the San Andreas fault after the 8 July 1986 North Palm-Springs earthquake. *Bull. Seismol. Soc. Am.* **78,** 1112–1122 (1988).
32. Bilham, R. Surface slip subsequent to the 24 November 1987 Superstition Hills, California, earthquake monitored by digital creepmeters. *Bull. Seismol. Soc. Am.* **79,** 424–450 (1989).
33. Sandwell, D., Sichoix, L., Agnew, D., Bock, Y. & Minster, J.-B. Near real-time radar interferometry of the M_w 7.1 Hector Mine Earthquake. *Geophys. Res. Lett.* **27,** 3101–3104 (2000).
34. Sharp, R., Rymer, M. & Lienkaemper, J. Surface displacement on the Imperial and Superstition Hills faults triggered by the Westmoreland, California, earthquake of 26 April 1981. *Bull. Seismol. Soc. Am.* **76,** 949–965 (1986).
35. Bürgmann, R. *et al.* Earthquake potential along the northern Hayward fault, California. *Science* **289,** 1178–1182 (2000).
36. Peltzer, G., Crampe, F., Hensley, S. & Rosen, P. Transient strain accumulation and fault interaction in the Eastern California shear zone. *Geology* **29,** 975–978 (2001).
37. Lettis, W., Wells, D. & Baldwin, J. Empirical observations regarding reverse earthquakes, blind thrust faults, and quaternary deformation: Are blind thrust faults truly blind? *Bull. Seismol. Soc. Am.* **87,** 1171–1198 (1997).
38. Dolan, J., Christofferson, S. & Shaw, J. Recognition of paleoearthquakes on the Puente Hills blind thrust fault, California. *Science* **300,** 115–118 (2003).
39. Sylvester, A. Strike-slip faults. *Geol. Soc. Am. Bull.* **100,** 1666–1703 (1988).
40. Ben-Zion, Y. *et al.* A shallow fault-zone structure illuminated by trapped waves in the Karadere-Duzce branch of the North Anatolian Fault, western Turkey. *Geophys. J. Int.* **152,** 699–717 (2003).
41. Fialko, Y. *et al.* Deformation on nearby faults induced by the 1999 Hector Mine earthquake. *Science* **297,** 1858–1862 (2002).
42. Hickman, S. & Zoback, M. Stress orientations and magnitudes in the SAFOD pilot hole. *Geophys. Res. Lett.* **31,** L15S12, doi:10.1029/2004GL20043 (2004).
43. Sieh, K. E. & Jahns, R. H. Holocene activity of the San Andreas fault at Wallace Creek, California. *Geol. Soc. Am. Bull.* **95,** 883–896 (1984).
44. Oskin, M. & Iriondo, A. Large-magnitude transient strain accumulation on the Blackwater fault, Eastern California shear zone. *Geology* **32,** 313–316 (2004).
45. Rosen, P., Hensley, S., Peltzer, G. & Simons, M. Updated repeat orbit interferometry package released. *Eos* **85,** 47 (2003).
46. Scharoo, R. & Visser, P. Precise orbit determination and gravity field improvement for the ERS satellites. *J. Geophys. Res.* **103,** 8113–8127 (1998).
47. Farr, T. & Kobrick, M. Shuttle Radar Topography Mission produces a wealth of data. *AGU Eos* **81,** 583–585 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Bürgmann for comments. This work was supported by the National Science Foundation and the Southern California Earthquake Center. Original Envisat ASAR data are copyright of the European Space Agency, acquired under CAT-1 research category. Aftershock locations were provided by M. Tatar, and coordinates of the geologically mapped faults by M. Heydari. P.R. conducted his work at JPL/Caltech, under contract with NASA.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.F. (fialko@radar.ucsd.edu).

© 2005 Nature Publishing Group

ARTICLES

Lack of long-term cortical reorganization after macaque retinal lesions

Stelios M. Smirnakis[1,2], Alyssa A. Brewer[3], Michael C. Schmid[1], Andreas S. Tolias[1], Almut Schüz[1], Mark Augath[1], Werner Inhoffen[4], Brian A. Wandell[3] & Nikos K. Logothetis[1]

Several aspects of cortical organization are thought to remain plastic into adulthood, allowing cortical sensorimotor maps to be modified continuously by experience. This dynamic nature of cortical circuitry is important for learning, as well as for repair after injury to the nervous system. Electrophysiology studies suggest that adult macaque primary visual cortex (V1) undergoes large-scale reorganization within a few months after retinal lesioning, but this issue has not been conclusively settled. Here we applied the technique of functional magnetic resonance imaging (fMRI) to detect changes in the cortical topography of macaque area V1 after binocular retinal lesions. fMRI allows non-invasive, *in vivo*, long-term monitoring of cortical activity with a wide field of view, sampling signals from multiple neurons per unit cortical area. We show that, in contrast with previous studies, adult macaque V1 does not approach normal responsivity during 7.5 months of follow-up after retinal lesions, and its topography does not change. Electrophysiology experiments corroborated the fMRI results. This indicates that adult macaque V1 has limited potential for reorganization in the months following retinal injury.

Under many conditions, the brain responds to the loss or distortion of input signals by adjusting the relevant cortical neuronal circuits. Treating dysfunction caused by peripheral damage depends crucially on understanding the nature of this cortical reorganization. Electrophysiological studies indicate that macaque primary visual cortex (V1) might retain a remarkable degree of plasticity into adulthood[1–5]. At 2–6 months after binocular retinal lesions that deprive a zone within V1 of its normal input, stimulus-driven activity is reported to recur up to 5 mm inside the border of the deafferented zone[1,2,4–6] (long-term reorganization). More modest changes in cortical topography spanning 1–2 mm are thought to occur immediately (minutes to hours) after deafferentiation[1,7–9] (but see refs 10 and 11).

However, it has not been entirely settled whether significant long-term reorganization occurs in area V1 (refs 12–14). Horton *et al.* find that after monocular retinal lesions in adult macaques, cytochrome oxidase levels in the deafferented V1 zone remain 'severely depressed' for many months, even after enucleation of the remaining eye[13]. Using electrophysiological recordings, Murakami *et al.* found no evidence for topographic reorganization near the deafferented zone induced in macaque V1 by a monocular retinal lesion, although perceptual filling-in clearly occurred at the corresponding visual field scotoma[12].

To clarify the extent and time course of reorganization in the visual system, we studied signals in macaque V1 after binocular retinal lesions. The lesions were made with a 532-nm photocoagulation laser and were located to create a homonymous visual field scotoma 4–8° in diameter (Fig. 1a–c). These lesions deprived part of V1 of visual input from each eye (cyan region, Fig. 1d), which is thought to be favourable for maximal reorganization[6,8]. We refer to the region in area V1 that is deprived of retinal input as the lesion projection zone (LPZ), following ref. 15.

Using 4.7-T functional magnetic resonance imaging (fMRI) in the anaesthetized macaque[16,17], we obtained maps of the visually driven activity in the monkey operculum surrounding the LPZ across time. The activity along the LPZ boundary was monitored every 2–8 weeks, starting 2–3 h after lesioning and continuing for 18–30 weeks. Visually driven activity was assessed by using monocular stimulation to avoid errors associated with potential eye misalignment. We analysed cortical activity along monocular LPZ borders that were also borders for the binocular LPZ (dotted line in Fig. 1d), where reorganization is expected to be maximal[6,8]. We used two visual stimulation paradigms to map cortical activity. In the first stimulation paradigm (full-field) blocks of an 18.4° × 14.1° rotating polar checkerboard pattern alternated with blocks of uniform background illumination. This stimulus strongly and reliably activates early visual areas[16,18]. In the second stimulation paradigm (expanding rings), dynamic annular patterns slowly expanded from fovea to periphery in order to assess the visual field eccentricity maps[19]. We measured the strength of the visually driven signal using coherence[20], which is equal to the amplitude of the blood-oxygenation-level-dependent (BOLD) signal modulation at the frequency of stimulus presentation, divided by the square root of the power over a range of nearby frequencies (see Methods). The two visual stimulation paradigms yielded the same result with respect to the extent of reorganization.

LPZ size measured by fMRI remains constant over time

Figure 2a, b shows a coherence map obtained with the expanding-ring stimulus overlaid on a flattened representation of the right visual cortex of the macaque designated E02. There is a well-circumscribed confluent region of low coherence (green) in the right operculum corresponding to the LPZ. The responses in the LPZ have a coherence of less than 0.4, whereas the coherence in immediately adjacent cortex is substantially higher. This level of coherence therefore served as a natural criterion for defining the border between LPZ and normal cortex. It is about 1 s.d. higher than the noise level of

[1]Max Planck Institute for Biological Cybernetics, Spemannstrasse 38, D-72076 Tübingen, Germany. [2]Department of Neurology, Massachusetts General Hospital, 55 Fruit Street, Boston, Massachusetts 02114, USA. [3]Neuroscience Program and Department of Psychology, Stanford University, Stanford, California 94305, USA. [4]Department of Ophthalmology I, University of Tübingen, Schleichstrasse 12, D-72076 Tübingen, Germany.

© 2005 Nature Publishing Group

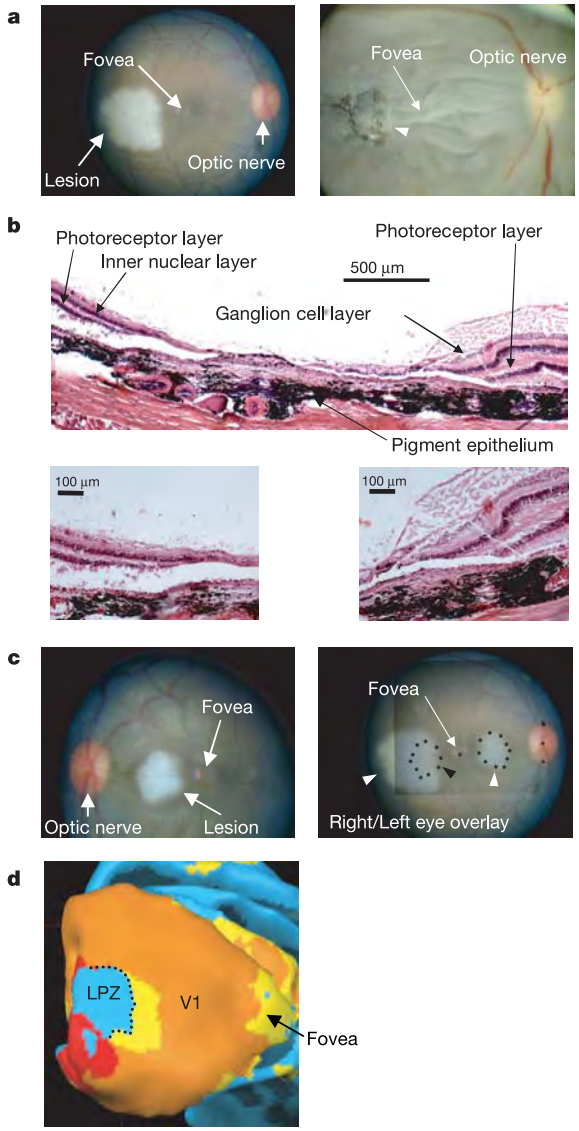


Figure 1 | Homonymous retinal lesions. a, Left, picture of the right eye fundus 1–2 h after photocoagulation. The lesion appears pale white. Right, right eye cup after extraction and fixation. The lesion (white arrowhead) is scarred and hyperpigmented. Its nasal border is at about 4° eccentricity and the temporal at about 12°. **b**, Top, haematoxylin–eosin stain of a section 15 μm thick through the right eye lesion in **a**. Note the complete destruction of the photoreceptor layer, as well as the nearly complete obliteration of ganglion and inner nuclear cell layers. Bottom left and right, edges of the haematoxylin-eosin stained section at higher magnification. The photoreceptor layer tapers near the lesion border zone (generally over 100–300 μm). **c**, Left, picture of the left eye fundus 1–2 h after photocoagulation. Right, overlay of the right and left fundi. Lesions are marked by white arrowheads. The left retina (smaller square overlay) was mirrored along the vertical axis and scaled to make the optic nerves overlap. Black dots outline the left eye lesion and its homonymous location in the right eye, which lies almost entirely over the right eye lesion except for a small sliver of normal retina (black arrowhead). This results in a partly homonymous left visual field scotoma. **d**, Eyes were stimulated separately with the full-field checkerboard stimulus (see Methods) and activation maps (voxels with $P < 0.05$, uncorrected, cluster criterion 6) were overlaid on the right operculum of monkey O97 by using Brain Voyager (Brain Innovation B.V., Maastricht, The Netherlands). Areas are coloured as follows: yellow, activation through the right eye; red, activation through the left eye; orange, activation map overlap; blue, no visual modulation through either eye. The dotted line marks the segment of the right-eye-LPZ border, where reorganization is expected to be maximal. This is the part of the right-eye-LPZ border that also belongs to the border of the binocular LPZ. The yellow region abutting the LPZ is the projection zone of the sliver of normal retina marked with the black arrowhead in (**c**, right).

coherence (0.28; see Methods). Using this criterion we outlined the LPZ on a flattened representation of the operculum at various time points: day 0 (black, LPZ_0), 4 months (cyan, LPZ_4) and 7.5 months (blue, LPZ_7) after lesioning (Fig. 2b). There was no significant change in the size of the LPZ over time.

To further ensure that this effect was not an artefact of the particular threshold or the coherence measure itself, we examined the signal modulation both within and adjacent to the LPZ. Figure 2d plots the average signal modulation inside LPZ_0 per stimulation cycle 7.5 months after lesioning; the corresponding plot in Fig. 2g shows the signal modulation derived from an area of approximately the same eccentricity located in normal cortex (see also Fig. 2f, h). We found no significant stimulus-driven modulation inside LPZ_0 even though 7.5 months had passed from the time of the lesion (Fig. 2c–e). Activity inside the V1 lesion projection zone does not return to anywhere near normal levels after homonymous binocular retinal lesions. This result holds true for both stimulation paradigms and for all four macaques used in this study (see also Supplementary Figs 1, 4 and 5). Small, rapid changes that stabilize before the end of the first fMRI session (up to 6 h after lesioning) cannot be excluded.

We also measured the profile of coherence as a function of distance perpendicular to the LPZ border and monitored its changes over time (Fig. 3 and Supplementary Fig. 2). This approach is sensitive and has the advantage that it does not depend on a particular choice of threshold. Using the anatomical scan we selected several thin (0.5–1.0 mm) regions of interest (ROIs) parallel to and spanning the BOLD LPZ border (Fig. 3a). The selected ROIs constituted a time-invariant frame of reference based on anatomy, relative to which we could measure changes in the profile of functional activity over time. We then plotted the mean coherence in each ROI as a function of the average distance of the ROI's voxels from the line of zero displacement (represented by the dotted lines in Fig. 3b, c). The line of zero displacement is a line parallel to the LPZ border that serves as the reference point for calculating distances. Although it can be thought of as an estimate of the LPZ border, its position can be chosen arbitrarily because the comparison of coherence profiles across time depends on relative distances only. For clarity we note here that it does not necessarily correspond to the criterion we used to define the LPZ border in Fig. 2b.

The coherence starts high in normal cortex outside the LPZ and decreases gradually to the noise-level coherence deep inside the LPZ, approximating a sigmoidal curve (Fig. 3b). The position and slope of this curve measure the position, the decline in functional activity and the spatial blurring of the signal from normal V1 into the LPZ. Cortical reorganization should be manifested by a shift of the curve towards the interior of the LPZ and/or by a decrease in its slope over time. Electrophysiology literature[1,4,6] suggests that shifts on the order of 5 mm could be expected 2–6 months after lesioning. Given BOLD's sensitivity[17,21,22] and our high ($1 \times 1\ mm^2$) in-plane functional resolution, we expect to be able to reliably detect border shifts greater than 1 mm, which is half an order of magnitude smaller than shifts reported in the electrophysiology literature[1,4,6]. In Fig. 3b we plot normalized coherence versus distance curves obtained with the full-field stimulus from monkey E02. Different time points are colour coded and overlaid on each other. No shift in position or decline in slope is apparent over the course of 7.5 months. Figure 3c summarizes the results for all monkeys for both retinotopic and full-field stimuli. The LPZ border remains stable to within 1 mm from its original location in all panels, and there is no systematic change in slope (Supplementary Fig. 2). These results remain true for the raw (un-normalized) coherence profiles (Supplementary Fig. 2) and are therefore not an artefact of coherence normalization.

Electrophysiological recordings across the LPZ border

A comparison between studies of contrast sensitivity that use the BOLD signal[17,21,22] and electrophysiological recordings[23] indicates that the BOLD signal is sensitive to small increases in multi-unit

© 2005 Nature Publishing Group

activity (smaller than 5% of the response elicited by the high-contrast checkerboard stimulus). However, the BOLD signal also reflects changes in local field potential underlying synaptic activity[17] that can be permanently disrupted inside the LPZ after deafferentiation. We therefore sought to corroborate the fMRI results presented above with electrophysiology experiments.

We performed three multi-electrode electrophysiology experiments after the conclusion of the fMRI sessions. We used a 16-electrode linear array (1-mm spacing) to record across the border of the LPZ (Fig. 4a) in monkey E02. Experimental conditions were identical to those in the fMRI experiments. We manually mapped V1 multi-unit receptive fields for two depths at each electrode location by using small oriented bars and square rotating checkerboard stimuli (Methods). Receptive field maps obtained from the same electrode at different cortical depths were similar in all cases. Figure 4b illustrates receptive field maps obtained in two such experiments. We mapped receptive fields starting with the electrode closest to the lunate (electrode no. 1) and proceeding towards the border of the LPZ. In all three experiments, our ability to identify multi-unit receptive fields ended at 1.6, 2.2 and 0.8 mm, respectively, outside the LPZ border defined using the fMRI data. We were unable to map classical receptive fields from any electrode located inside the BOLD-defined LPZ border (0/52 recording sites, including depth).

Figure 4c shows firing rate histograms obtained during 10 s of full-field stimulation alternated with 10 s of uniform illumination. During stimulus ON periods, electrode locations outside the LPZ (electrodes 2–10) have steady-state firing rates significantly higher than the spontaneous rates achieved during uniform illumination. In

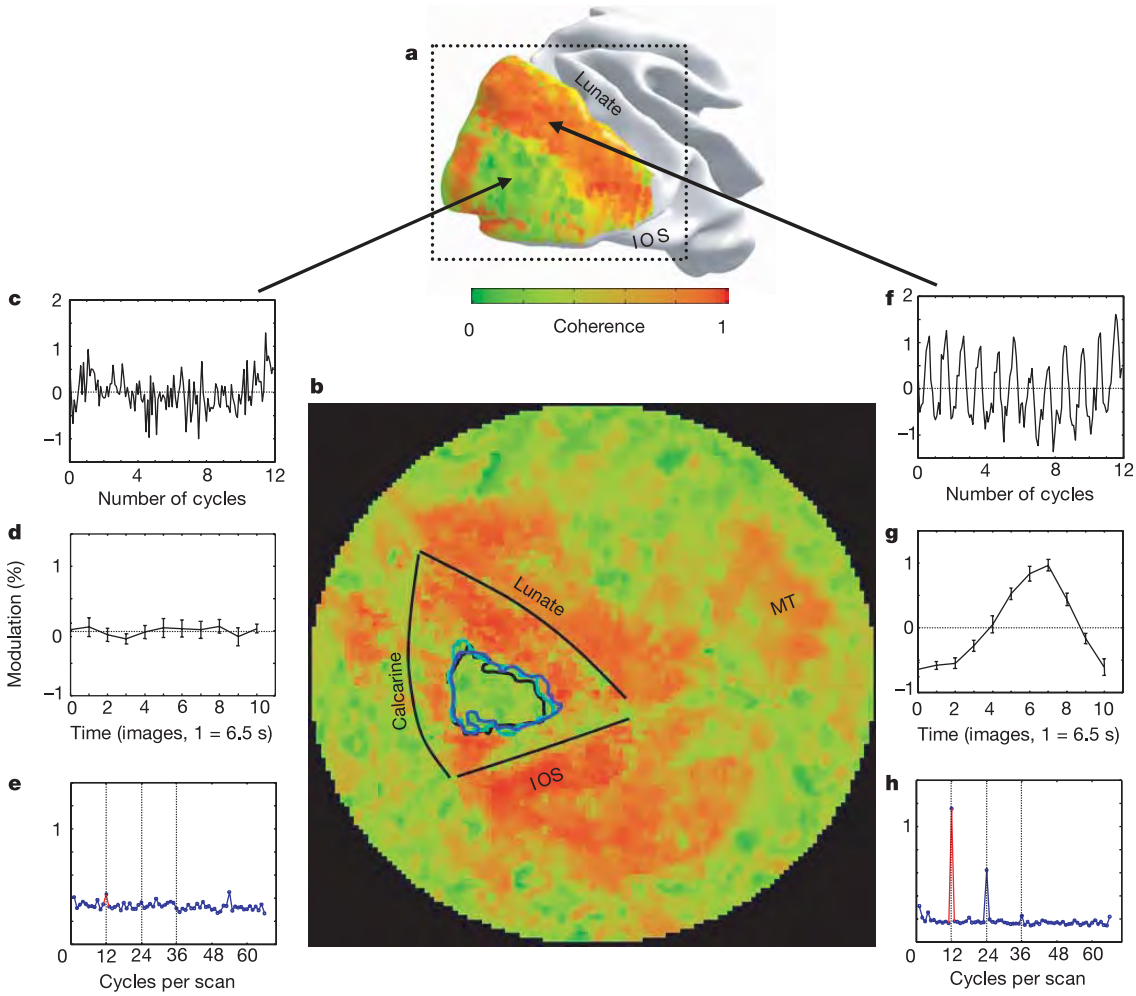


Figure 2 | BOLD signal inside the LPZ remains at noise level and LPZ size does not change as a function of time after lesioning. **a**, Coherence map on the right operculum measured with the expanding-ring stimulus (see Methods) presented to the right eye of macaque E02 4 months after lesioning. The green area is the LPZ, a zone of low coherence surrounded by normal cortex. Note that noise level coherence is expected to be 0.28 and not zero (see Methods). **b**, A cortical area of radius 3 cm centred near the fovea was flattened and the coherence map was overlaid. The monkey operculum is outlined by the calcarine, lunate and inferior occipital sulci. Regions of low coherence outside the operculum correspond to non-visual cortex or visual cortex too eccentric to be activated by our stimuli. LPZ borders are shown at 0 days (LPZ_0, black), 4 months (LPZ_4, cyan) and 7.5 months (LPZ_7, blue) after lesioning. LPZ size was, respectively, 158, 179 and 180 mm^2 and was larger, if anything, at later time points. Because the LPZ perimeter is about 50 mm this corresponds approximately to an 0.5-mm outward shift of the LPZ border (within noise). A threshold coherence level of 0.4, 1 s.d. above noise level, was used to define the LPZ border at all time points. **c**–**e**, Signal in LPZ_0 recorded 7.5 months after lesioning. **c**, Percentage modulation of the BOLD signal as a function of stimulus cycles. **d**, Average signal modulation per stimulus cycle (11 rings). **e**, Signal amplitude as a function of temporal frequency. Red marks the frequency of the visual stimulus. Note that the signal amplitude at the stimulus frequency does not differ significantly from the amplitude of noise present at other frequencies ($P > 0.05$). There is therefore no significant stimulus-driven modulation inside the LPZ_0 even 7.5 months after lesioning. **f**–**h**, Plots corresponding to **c**–**e** for a region of normal V1 cortex at similar eccentricity to that of the LPZ. Strong visual modulation is evident in all plots. IOS, inferior occipital sulcus; MT, middle temporal area. Error bars represent s.e.m.

© 2005 Nature Publishing Group

contrast, the steady-state firing rates of electrodes inside the LPZ during stimulus ON periods are similar to their firing rates during uniform illumination (Fig. 4c). In Fig. 4d we plot the firing rate modulation strength at steady state during ON/OFF stimulation as a function of distance from the BOLD-defined LPZ border, by pooling and aligning appropriately data from all three experiments. On average, cortical locations inside the LPZ are not significantly modulated by the stimulus at steady state (mean modulation 1.3% with 95% confidence interval −2.7 to 5.3%). The border of the LPZ defined by electrophysiology (blue curve) lies about 1 mm outside the BOLD-defined LPZ border (red curve). The steady-state electrophysiological measurements therefore agree well with the BOLD-defined LPZ border.

The neurons in the LPZ were not entirely silent. There were transient responses at both stimulus onset and offset (Fig. 4c, electrodes 11–16). These transients were insufficient to cause a BOLD response inside the LPZ, probably because they were brief (generally less than 1 s) compared with the 6.0–6.5 s required for the acquisition of one whole-brain image. It is possible that such transients reflect a degree of cortical reorganization, but we do not

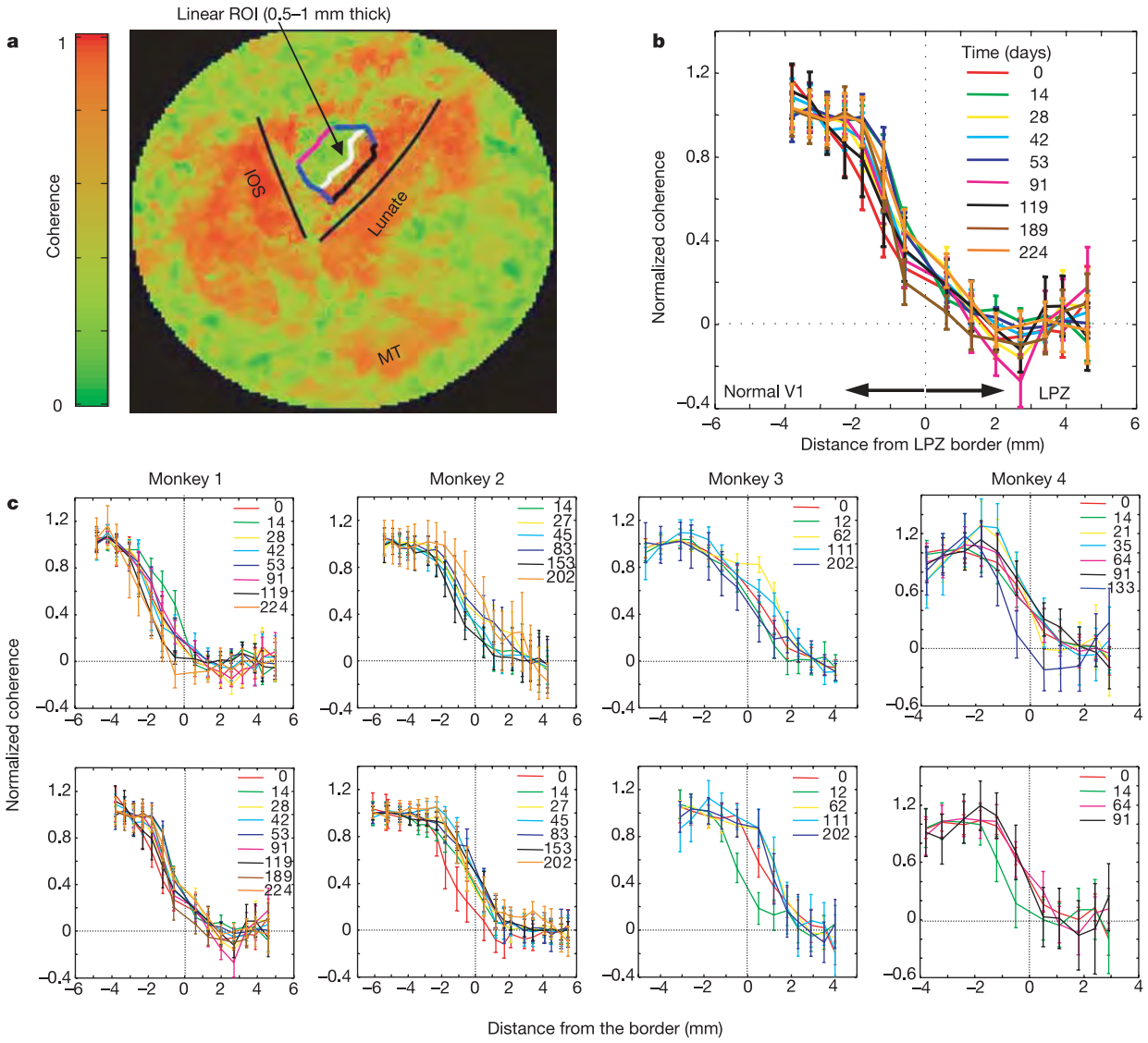


Figure 3 | The profile of BOLD coherence across the LPZ border does not change as a function of time after lesioning. **a**, Flattened map of the right operculum of macaque E02 with coherence map overlaid, as in Fig. 2. A set of linear ROIs 0.5–1.0-mm thick, parallel to the LPZ border, was selected to span several millimetres on each side of the border. The white line represents the first ROI located inside the scotoma. The purple line denotes the last ROI selected inside the scotoma, the black line the last ROI selected outside. The blue outline represents the endpoints of all ROIs selected. ROIs are selected once on the anatomical scan and remain the same for all time points. They therefore provide an absolute, time-invariant, reference frame that can be used to gauge whether the profile of functional activity shifts over time. **b**, Monkey E02, full-field stimulation paradigm. The mean coherence of each linear ROI is plotted against the mean distance of that ROI's voxels from the line of zero displacement (represented by the dotted lines at point zero). The resulting sigmoidal curves were normalized to reach a plateau at 0 and 1 by subtracting the mean of the last four points (that is, noise coherence deep inside the LPZ) and then dividing by the mean of the first four points (that is, the strength of visually driven modulation in normal cortex). Before performing the subtractive normalization we confirmed that the average coherence over the last four points of the sigmoidal curve (which lie inside the LPZ) remained at noise levels throughout. Time is measured in days after lesioning. Different time points are colour coded. Positive distances are towards the LPZ interior, negative towards normal cortex. **c**, Equivalent plots to those in **b** for each monkey (from left to right: E02, A01, O97 and E01). Top row, retinotopic stimulation paradigm; bottom row, full-field stimulation paradigm. Error bars represent s.e.m. There is no consistent shift in position or change in slope of the coherence–distance curve over time (see also Supplementary Fig. 2). The position of the LPZ border is stable to within 1 mm. IOS, inferior occipital sulcus; MT, middle temporal area.

© 2005 Nature Publishing Group

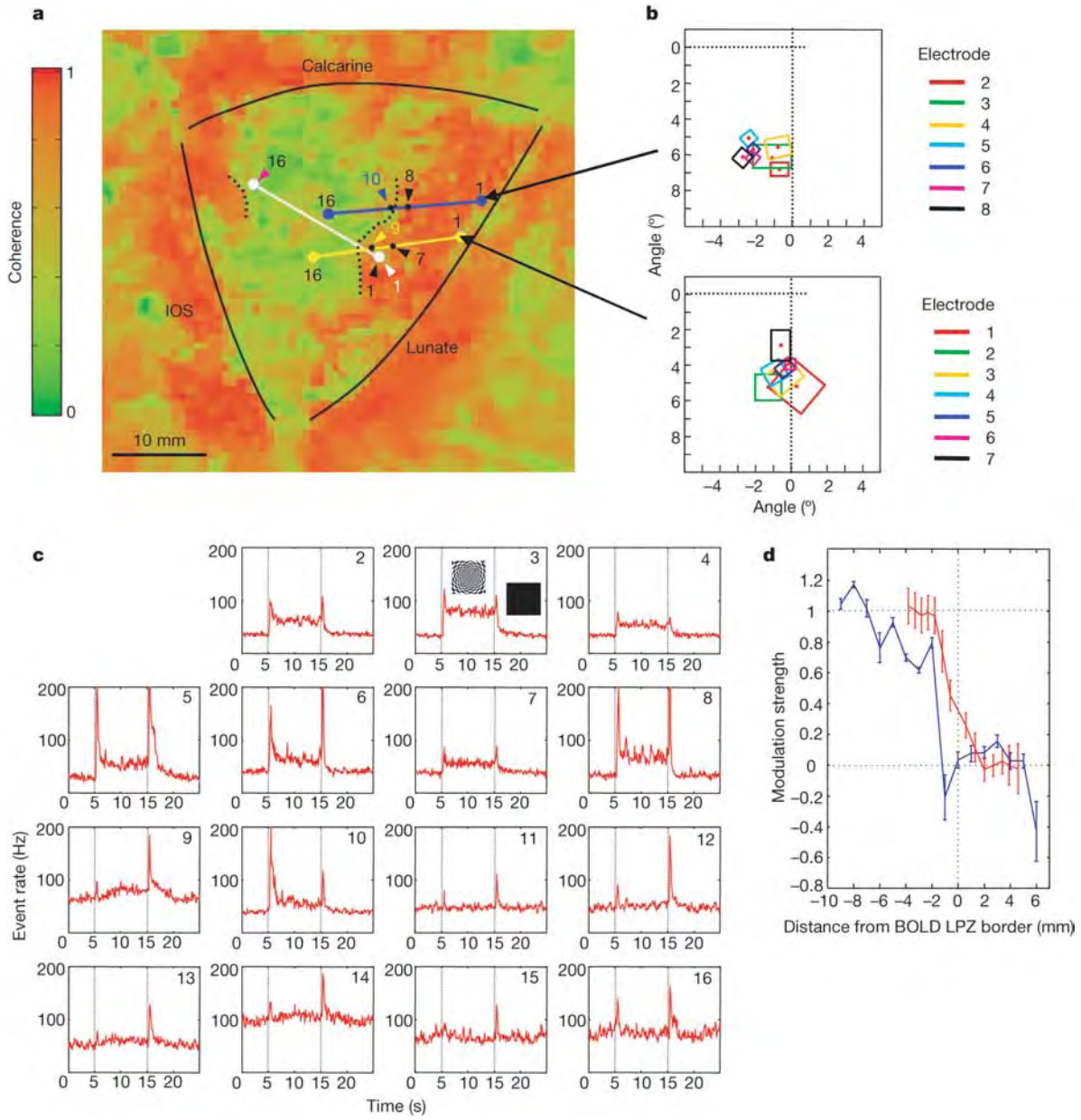


Figure 4 | Receptive field maps cannot be obtained and steady-state neural responses remain abnormal inside the LPZ. a, Coherence map (day 0, full-field stimulus) overlaid on the flattened right operculum of monkey E02. Each linear trajectory (white, yellow and blue) represents 16 recording electrode locations from one of three experiments. Electrodes were 1 mm apart, starting near the lunate (electrode 1). Black arrowheads mark the last position yielding a multi-unit RF. Blue, yellow and white arrowheads mark the last position whose steady-state multi-unit activity was modulated more than 10% by the full-field stimulus. Dotted lines identify segments of the BOLD-defined LPZ border. The dotted line on the right corresponds to point 0 in Fig. 3b. **b**, Multi-unit RF maps from experiments 2 and 3. RFs could not be mapped with spatially restricted stimuli beyond electrode 1 (experiment 1, data not shown), 7 (experiment 2) and 8 (experiment 3). RF location is accurate to 1–2° because of slow eye drifts. **c**, Multi-unit firing rate histograms binned at 100 ms for electrodes in experiment 3 (numbers at top right), calculated by averaging 48 cycles of full-field visual stimulation alternating with spatially uniform background intensity (electrode 3 inset). Dotted lines mark stimulus ON (5 s) and OFF (15 s) transitions. Multi-unit activity at steady state is higher during stimulation for electrodes 2–10 (outside the LPZ), but not for electrodes 11–16 (inside the LPZ). Electrode 10 falls on the BOLD-defined LPZ border. Brief ON and OFF firing transients persist inside the LPZ (electrodes 11–16). These may arise through extra-classical receptive field effects known to exist in normal V1 (see Discussion and Supplementary Fig. 3). Transients arising in electrodes with good RF maps (electrodes 2–8) had a mean latency of 68 ± 4 ms compared with 93 ± 16 ms for electrodes near or inside the LPZ border (electrodes 9–16). Electrode 1 was damaged on entry. **d**, Steady-state firing rate (blue) and coherence (red) modulation curves as a function of distance from the BOLD-defined LPZ border. The event-rate modulation strength was calculated as the steady-state (discarding the first 2 s after each transition) firing rate with stimulus ON minus OFF, divided by ON and normalized by dividing by the mean of the first four points, located in normal cortex. Data from all three physiology experiments were used after being aligned appropriately with respect to the LPZ border. Inside the LPZ the average event-rate modulation strength was 1.3%, with 95% confidence limits from −2.7 to 5.3%. The 224th day coherence–distance curve (red) from Fig. 3b is plotted for comparison. Accuracy along the *x* axis is within 1 mm. Error bars represent s.e.m. IOS, inferior occipital sulcus.

© 2005 Nature Publishing Group

think this is likely because we did not observe these transients inside the LPZ during receptive field mapping with spatially restricted stimuli. This indicates that the transients might arise through the spatial summation of extra-classical receptive field responses[24–27] that are present in normal macaque V1 and can extend over considerable cortical distances[25,28]. Indeed, we observed such transients in animals without retinal lesions by stimulating the extra-classical surround (see Supplementary Fig. 3). It is possible that these transient responses are accentuated to somewhat above their normal level by the downregulation of inhibition that is thought to occur inside the LPZ[29,30]. However, even if some plasticity is mediated through these transient responses, it is clear that V1 cortex has not achieved anything like normal responsivity: receptive field maps cannot be obtained and there is no sustained firing to 3.5-Hz visual modulation inside the LPZ border.

Discussion

Our results contrast with studies[1,3,4,6,31] indicating that substantial long-term reorganization occurs over 2–6 months and that the LPZ border shifts by about 5 mm (refs 1,4). There could be several reasons for this discrepancy. First, long-term reorganization might be restricted to a subset of few sparsely distributed single neurons that form the basis of the physiology experiments. The BOLD response reflects the average activation of ensembles of cells and thus might provide a more complete assessment of overall neuronal recovery. It is also less prone to possible selection bias, which might be a consideration for physiology experiments. Second, changes in the apparent size of the LPZ might occur without reorganization because of retinal recovery from the swelling caused by photocoagulation[13]. Most studies of long-term reorganization are susceptible to this effect because they compare cortical activity immediately after retinal photocoagulation with activity seen several months later[1,3,4]. The width of the LPZ border as judged by the sloping segment of the sigmoidal plots presented in Fig. 3 is about 3–4 mm, corresponding to 1–2° of visual angle or a retinal distance of 150–300 μm. This is similar in extent to the partial damage to the photoreceptor layer seen at the edge of the lesion (Fig. 1b). Recovery of this area, which might be 'stunned' immediately after photocoagulation, could lead to apparent activity shifts of 3–4 mm. In our experiments, we did not observe a significant shift of the LPZ border from the day of the lesion onwards, so it is unlikely that retinal 'stunning' had a significant role. Last, estimates of long-term recovery might be exaggerated by a systematic underestimation, perhaps resulting from retinal 'stunning' of the extent of immediate reorganization (minutes to hours). Several studies[1,7–10,32,33] report immediate receptive field expansion after stimulus deprivation or retinal lesioning (but see ref. 11). Immediate changes extending more than 4.5 mm from the LPZ border have been reported[10] in cat area 17 after retinal detachment. Such changes[9,10] are commensurate with the extent of long-term reorganization reported elsewhere[1,4,6], making it difficult to decipher what, if any, component of reorganization is long-term. Previous reports indicate that a significant fraction (about 10%) of parafoveal (2–5°) macaque V1 neurons receive excitatory inputs from 2° or further away[28,34]. It is then possible that the reported receptive field shifts reflect activity arising in the far reaches of the classical receptive field of cortical cells. This activity may initially be subthreshold, then rise above threshold after lesioning through relatively fast adaptive gain adjustments or downregulation of inhibition[29,30]. Retinal 'stunning' in the lesion surround might prevent this from manifesting immediately, causing it to masquerade later as 'long-term' cortical reorganization.

Our results also contrast with a recent report by Baker *et al.* of extensive reorganization in human visual areas, including area V1 (ref. 35). Those measurements were made in two adult human subjects who had been diagnosed with macular degeneration covering the fovea nearly 20 years earlier. In contrast, Sunness and Yantis found no significant plasticity in a human subject with a partial macular lesion diagnosed 2 years before measurement[36], and this has been confirmed in a second subject with a relatively complete macular lesion (B.A.W., J. Liu and S. Nagadomari, unpublished observations). These human measurements, coupled with the present macaque data, indicate that large-scale cortical reorganization in adult V1 might not occur until at least several years after sensory deprivation.

Under the conditions of our experiment, intra-cortical horizontal connections are apparently not sufficient to allow V1 to regain anywhere near its previous functionality after retinal lesions, even though they might be sprouting new axonal boutons[37]. Should horizontal connections also prove ineffective elsewhere, this could support the notion that large-scale reorganization reported in other areas[38,39] (as well as in the visual system in early blindness[40–42]) might be mediated subcortically or through feedback projections[43,44]. In general, connections that span large distances do not easily form *de novo* in the adult central nervous system[45,46], so plasticity is unlikely to be manifested in cases in which the formation of such connections is required. This is more likely to be true in areas with spatially restricted connectivity patterns and small, largely non-overlapping receptive fields (like V1). In contrast, areas with large, overlapping receptive fields and broad connectivity patterns might be able to reorganize effectively without requiring the formation of an extensive network of new connections (Fig. 5).

In summary, we have measured the spatial pattern of visually driven activity near the border of the lesion projection zone created in adult macaque V1 by a homonymous retinal lesion. Over 4–7 months we observed no significant change in the position of the BOLD-defined LPZ border. BOLD responses in the LPZ did not become visually driven, and this was confirmed by electrophysiological measurements of multi-unit action potentials. Taken together, these experiments indicate that little long-term reorganization

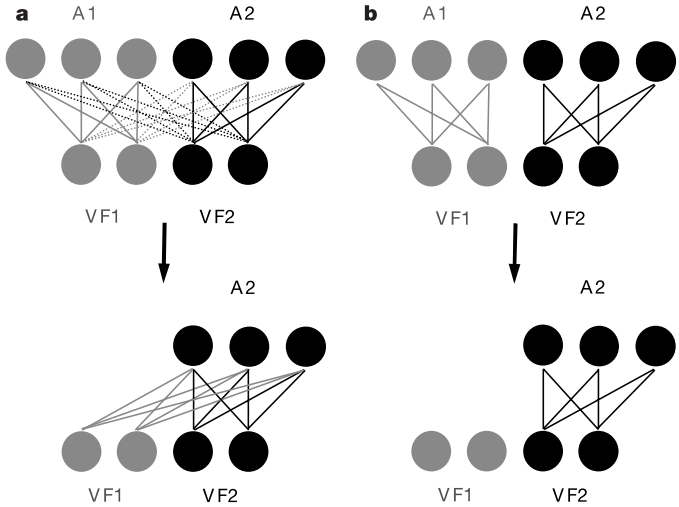


Figure 5 | Examples of connectivity diagrams. a, The top layer of nodes is the union of two subregions A1 (grey) and A2 (black) with overlapping receptive fields but with connections that have different strengths (dotted, weak; solid, strong) so that A1 processes primarily input in the VF1 sector of the visual field and A2 in the VF2 sector. Although injury to area A1 will create an immediate processing deficit in sector VF1, over time the existing weak connections from A2 to VF1 (dotted grey lines) may strengthen (bottom, solid grey lines) and compensate. This is akin to the robustness seen in densely connected neural networks that have a portion of their nodes damaged. **b**, Such a scheme is not possible if the connectivity pattern is restricted and cells have non-overlapping receptive fields. For area A2 to assume processing in sector VF1, new connections have to form *de novo*, and this may be difficult in the adult central nervous system. Pharmacological manipulations to enhance this process hold great promise for enhancing plasticity in the central nervous system in the future.

© 2005 Nature Publishing Group

occurs in the LPZ and that neuronal responses in the LPZ do not recover to anything approaching their normal state. On the basis of these results a more critical examination of previous ideas about the degree and temporal course of area V1 reorganization after retinal lesions might be worthwhile.

Beyond the specific study of V1, we showed that fMRI could be used effectively to monitor cortical organization in the anaesthetized macaque preparation. The use of macaque fMRI in combination with experimental manipulations that can enhance plasticity is a powerful approach that promises to enhance our understanding of the processes of neural recovery and reorganization, and to provide a useful model for studying human recovery from stroke.

METHODS

Retinal lesions. Using a photocoagulation laser (NIDEK GYC-2000; 532 nm) we induced homonymous retinal lesions in adult rhesus macaques under general anaesthesia (Fig. 1). The lesions required 10–77 overlapping small laser burns (power 310 mW, duration 0.2 s each) to obtain a single, round, well demarcated, homogeneous, extrafoveal scar with a diameter ranging from 600 to 1,200 μm (Fig. 1). This destroyed the photoreceptor and bipolar cell layers as well as most of the ganglion cell layer (Fig. 1b). Histological confirmation was obtained in two of four animals (the remaining two are still alive). These animals were killed with pentobarbital under general anaesthesia, and the eyes were extracted and fixed by immersion in 3.7% formaldehyde prepared in 0.1 M sodium phosphate buffer. The eyes were later embedded and processed in Paraplast; horizontal sections 15–20-μm thick were cut and stained with haematoxylin and eosin. Animal fundi were photographed in each experimental session, and we confirmed that all lesions remained stable throughout the course of the experiments.

MRI parameters. We used multi-shot (eight-segment) gradient-recalled echo-planar imaging in a BIOSPEC 4.7 T/40 cm scanner (Bruker) with 50 mT m^{-1} gradients. The acquisition parameters were TE = 20 ms, TR = 750 or 805 ms, flip angle = 40°. Either 15 or 17 axial slices were collected at 1 × 1 mm^2 in-plane resolution and 2-mm thickness. These were resampled accordingly and aligned to a full-brain anatomical scan acquired before lesioning by using an MDEFT sequence at 0.5 × 0.5 × 0.5 mm^3 resolution. ROIs were defined on the flattened version of each monkey's anatomical scan, and had an anatomical resolution of about 0.5 × 0.5 mm^2 along the cortical sheet.

Animal preparation. Four healthy adult *Macaca mulatta*, older than 4 years, were used for the experiments. All sessions were in full compliance with the guidelines of the European Community for the care and use of the laboratory animals (EUVD 86/609/EEC) and were approved by the local authorities (Regierungspräsidium). fMRI experiments were performed under general anaesthesia in accordance with previously published protocols[16,19]. The animal was intubated after induction with fentanyl (31 μg kg^{-1}), thiopental (5 mg kg^{-1}) and succinylcholine chloride (3 mg kg^{-1}); anaesthesia was maintained with isoflurane (animals O97 and E01) or remifentanyl (0.5–2 μg kg^{-1} min^{-1}; animals E02 and A01). Mivacurium chloride (5–7 mg kg^{-1} h^{-1}) was used after induction to ensure complete paralysis of the eye muscles. Lactate Ringer's with 2.5% glucose was infused at 10 ml kg^{-1} h^{-1} throughout the experiment. The temperature was maintained at 38–39 °C. After achieving mydriasis with two drops of 1% cyclopentolate hydrochloride, each eye was fitted with hard contact lenses (Harte PMMA-Linsen; Wöhlk, Kiel, Germany) to bring it to focus on the stimulus plane (about 2 dioptres).

Visual stimulation. Stimuli were presented monocularly, always in the same eye for each animal, with an SVGA fibre-optic system (AVOTEC; Silent Vision) with a resolution of 640 pixels × 480 pixels and a 60-Hz frame rate.

In the first stimulation model (full field) we presented blocks of a 18.4° × 14.1° rotating polar checkerboard pattern (corresponding to 3.5-Hz visual modulation at 100% contrast) alternating with blocks of uniform background illumination every 30 or 32 s, depending on the animal, for a total of 12 cycles. For animals O97 and E01, 48-s blocks for four cycles were also used. This stimulus has been shown[16,18] to activate powerfully early visual areas. At times a 3.7° diameter dark circular occluder centred at 3.7° in the visual field opposite to the scotoma was also presented for control purposes.

In the second stimulation model (expanding rings), we used annuli 0.9° wide[19] whose outer radius expanded from either 0.3° or 0.9° to 6.9° in steps of 0.6° (12 or 11 annuli). The annular patterns were rotating checkerboards. Each annulus was presented for 6.0 or 6.5 s and the entire sequence of annuli was repeated 12 or 18 times per scan, depending on the animal.

Analysis. We used the mrVista software (Stanford, http://white.stanford.edu/software) for data analysis[19,47]. We use the coherence of the fMRI series at the fundamental stimulus frequency to measure the strength of the BOLD responses[19,20]. Our measure for coherence is

$$C = A(f_0) \Big/ \left(\sum_{f_0-\Delta f/2}^{f_0+\Delta f/2} A(f)^2 \right)^{1/2}$$

where f_0 is the stimulus presentation frequency, $A(f_0)$ the amplitude of the signal at that stimulation frequency, and Δf a range of frequencies around the fundamental. For the stimuli that we used, f_0 corresponded to 12 or 18 cycles per scan and Δf was chosen to be a bandwidth of 12 cycles centred at f_0. For sessions with f_0 equal to four cycles we set Δf from three to seven cycles. Note that the noise level of coherence depends on the chosen bandwidth so that, for $\Delta f = 12$ cycles, coherence in the absence of visual modulation is expected to be $1/(\Delta f + 1)^{0.5} = 0.28$.

Electrophysiology. The animal (E02) was anaesthetized with remifentanyl and paralysed as described above for the fMRI experiments. Visual stimuli were presented monocularly through a high-resolution (1,024 pixels × 768 pixels) AVOTEC device. A linear array of 16 tungsten electrodes (impedance 200–800 kΩ) at 1-mm intervals was placed across the border of the scotoma through a craniotomy that exposed the dura. The electrodes were lowered one by one under audio guidance and were advanced about 300 μm from the point at which multi-unit activity was first audible. They were then allowed to settle for 15–20 min before manual receptive field (RF) mapping. RFs were mapped by using oriented bars (sizes 0.5° × 1°, 0.5° × 2°, 1° × 2° and 1° × 3°, with orientations 0°, 45°, 90° and 135°) first moved for a manual sweep of the central 12° of the monkey's left visual field, then flashed at 0.5–2.0 Hz over a range of promising visual field locations. Flashed and rotating square polar checkerboard stimuli (sizes 1° × 1°, 2° × 2° and 3° × 3°) were also used. The pattern of the sweep was standardized to ensure that we did not skip any receptive field locations, and care was taken to define the border of each multi-unit receptive field approximated by a rectangle. All the stimuli described were used for each experiment except the first, for which only the polars were used. At the end of RF mapping we advanced the electrode another 0.5–1.0 mm, waited for 15 min and repeated the process. In all penetrations, superficial and deep receptive fields were approximately commensurate and at the same location. Eye drift was about 1–2° during an experiment.With all electrodes lowered we recorded multi-unit activity under stimulation with the rotating full-field polar checkerboard alternating with uniform background illumination (every 10 s for experiments 2 and 3, and every 5 s for experiment 1). The signal was digitized at 20,833 Hz, band-pass filtered to lie between 312 and 10 kHz, and processed with Data Visualization and Analysis Software (Plexon) to select spikes with an amplitude greater than 3 s.d. above the mean. Multi-unit firing rate histograms (Fig. 4c) were calculated by binning the spike trains in 100-ms bins and averaging together all stimulus cycles (48 for each experiment). Latency was computed as the time after stimulus transition when the firing rate was more than 10% of the transient's peak for three consecutive 5-ms time bins.

Received 2 December 2004; accepted 16 February 2005.

1. Gilbert, C. D. & Wiesel, T. N. Receptive field dynamics in adult primary visual cortex. *Nature* **356**, 150–152 (1992).
2. Gilbert, C. D., Hirsch, J. A. & Wiesel, T. N. Lateral interactions in visual cortex. *Cold Spring Harb. Symp. Quant. Biol.* **55**, 663–677 (1990).
3. Heinen, S. J. & Skavenski, A. A. Recovery of visual responses in foveal V1 neurons following bilateral foveal lesions in adult monkey. *Exp. Brain Res.* **83**, 670–674 (1991).
4. Darian-Smith, C. & Gilbert, C. D. Topographic reorganization in the striate cortex of the adult cat and monkey is cortically mediated. *J. Neurosci.* **15**, 1631–1647 (1995).
5. Gilbert, C. D. Adult cortical dynamics. *Physiol. Rev.* **78**, 467–485 (1998).
6. Kaas, J. H. *et al.* Reorganization of retinotopic cortical maps in adult mammals after lesions of the retina. *Science* **248**, 229–231 (1990).
7. Pettet, M. W. & Gilbert, C. D. Dynamic changes in receptive-field size in cat primary visual cortex. *Proc. Natl Acad. Sci. USA* **89**, 8366–8370 (1992).
8. Chino, Y. M., Kaas, J. H., Smith, E. L., Langston, A. L. & Cheng, H. Rapid reorganization of cortical maps in adult cats following restricted deafferentation in retina. *Vision Res.* **32**, 789–796 (1992).
9. Calford, M. B., Schmid, L. M. & Rosa, M. G. Monocular focal retinal lesions induce short-term topographic plasticity in adult cat visual cortex. *Proc. R. Soc. Lond. B* **266**, 499–507 (1999).
10. Schmid, L. M., Rosa, M. G. & Calford, M. B. Retinal detachment induces massive immediate reorganization in visual cortex. *Neuroreport* **6**, 1349–1353 (1995).
11. DeAngelis, G. C., Anzai, A., Ohzawa, I. & Freeman, R. D. Receptive field structure in the visual cortex: does selective stimulation induce plasticity? *Proc. Natl Acad. Sci. USA* **92**, 9682–9686 (1995).
12. Murakami, I., Komatsu, H. & Kinoshita, M. Perceptual filling-in at the scotoma

 © 2005 Nature Publishing Group

following a monocular retinal lesion in the monkey. *Vis. Neurosci.* **14,** 89–101 (1997).
13. Horton, J. C. & Hocking, D. R. Monocular core zones and binocular border strips in primate striate cortex revealed by the contrasting effects of enucleation, eyelid suture, and retinal laser lesions on cytochrome oxidase activity. *J. Neurosci.* **18,** 5433–5455 (1998).
14. Rosa, M. G., Schmid, L. M. & Calford, M. B. Responsiveness of cat area 17 after monocular inactivation: limitation of topographic plasticity in adult cortex. *J. Physiol. (Lond.)* **482,** 589–608 (1995).
15. Schmid, L. M., Rosa, M. G., Calford, M. B. & Ambler, J. S. Visuotopic reorganization in the primary visual cortex of adult cats following monocular and binocular retinal lesions. *Cerebral Cortex* **6,** 388–405 (1996).
16. Logothetis, N. K., Guggenberger, H., Peled, S. & Pauls, J. Functional imaging of the monkey brain. *Nature Neurosci.* **2,** 555–562 (1999).
17. Logothetis, N. K., Pauls, J., Augath, M., Trinath, T. & Oeltermann, A. Neurophysiological investigation of the basis of the fMRI signal. *Nature* **412,** 150–157 (2001).
18. Tolias, A. S., Smirnakis, S. M., Augath, M. A., Trinath, T. & Logothetis, N. K. Motion processing in the macaque: revisited with functional magnetic resonance imaging. *J. Neurosci.* **21,** 8594–8601 (2001).
19. Brewer, A. A., Press, W. A., Logothetis, N. K. & Wandell, B. A. Visual areas in macaque cortex measured using functional magnetic resonance imaging. *J. Neurosci.* **22,** 10416–10426 (2002).
20. Engel, S. A., Glover, G. H. & Wandell, B. A. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cerebral Cortex* **7,** 181–192 (1997).
21. Boynton, G. M., Engel, S. A., Glover, G. H. & Heeger, D. J. Linear systems analysis of functional magnetic resonance imaging in human V1. *J. Neurosci.* **16,** 4207–4221 (1996).
22. Wandell, B. A. Computational neuroimaging of human visual cortex. *Annu. Rev. Neurosci.* **22,** 145–173 (1999).
23. Sclar, G., Maunsell, J. H. R. & Lennie, P. Coding of image contrast in central visual pathways of the macaque monkey. *Vision Res.* **30,** 1–10 (1990).
24. Lee, T. S., Mumford, D., Romero, R. & Lamme, A. F. The role of the primary visual cortex in higher level vision. *Vision Res.* **38,** 2429–2454 (1998).
25. Rossi, A. F., Desimone, R. & Ungerleider, L. G. Contextual modulation in primary visual cortex of macaques. *J. Neurosci.* **21,** 1698–1709 (2001).
26. Li, W., Thier, P. & Wehrhahn, C. Neuronal responses from beyond the classic receptive field in V1 of alert monkeys. *Exp. Brain Res.* **139,** 359–371 (2001).
27. Rossi, A. F., Rittenhouse, C. D. & Paradiso, M. A. The representation of brightness in primary visual cortex. *Science* **273,** 1104–1107 (1996).
28. Cavanaugh, J. R., Bair, W. & Movshon, J. A. Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *J. Neurophysiol.* **88,** 2530–2546 (2002).
29. Hendry, S. H. & Jones, E. G. Reduction in number of immunostained GABAergic neurones in deprived-eye dominance columns of monkey area 17. *Nature* **320,** 750–756 (1986).
30. Rosier, A. M. *et al.* Effect of sensory deafferentiation on immunoreactivity of GABAergic cells and on GABA receptors in the adult cat visual cortex. *J. Comp. Neurol.* **359,** 476–489 (1995).
31. Das, A. & Gilbert, C. D. Long-range horizontal connections and their role in cortical reorganization revealed by optical recording of cat primary visual cortex. *Nature* **375,** 780–784 (1995).
32. Volchan, E. & Gilbert, C. D. Interocular transfer of receptive field expansion in cat visual cortex. *Vision Res.* **35,** 1–6 (1995).
33. Fiorani, M. Jr, Rosa, M. G. P., Gattass, R. & Rocha-Miranda, C. E. Dynamic surrounds of receptive fields in primate striate cortex: A physiological basis for perceptual completion? *Proc. Natl Acad. Sci. USA* **USA89,** 8547–8551 (1992).
34. Levitt, J. B. & Lund, J. S. The spatial extent over which neurons in macaque striate cortex pool visual signals. *Vis. Neurosci.* **19,** 439–452 (2002).
35. Baker, C. I., Peli, E., Knouf, N. & Kanwisher, N. G. Reorganization of visual processing in macular degeneration. *J. Neurosci.* **25,** 614–618 (2005).
36. Sunness, J. S., Liu, T. & Yantis, S. Retinotopic mapping of the visual cortex using functional magnetic resonance imaging in a patient with central scotomas from atrophic macular degeneration. *Ophthalmology* **111,** 1595–1598 (2004).
37. Darian-Smith, C. & Gilbert, C. D. Axonal sprouting accompanies functional reorganization in adult cat striate cortex. *Nature* **368,** 737–740 (1994).
38. Pons, T. P., Garraghty, P. E. & Mishkin, M. Lesion-induced plasticity in the second somatosensory cortex of adult macaques. *Proc. Natl Acad. Sci. USA* **85,** 5279–5281 (1988).
39. Pons, T. P. *et al.* Massive cortical reorganization after sensory deafferentation in adult macaques. *Science* **252,** 1857–1860 (1991).
40. Sadato, N., Okada, T., Honda, M. & Yonekura, Y. Critical period for cross-modal plasticity in blind humans: a functional MRI study. *Neuroimage* **16,** 389–400 (2002).
41. Sadato, N. *et al.* Activation of the primary visual cortex by Braille reading in blind subjects. *Nature* **380,** 526–528 (1996).
42. Cohen, L. G. *et al.* Functional relevance of cross-modal plasticity in blind humans. *Nature* **389,** 180–183 (1997).
43. Angelucci, A., Levitt, J. B. & Lund, J. S. Anatomical origins of the classical receptive field and modulatory surround field of single neurons in macaque visual cortical area V1. *Prog. Brain Res.* **136,** 373–388 (2002).
44. Angelucci, A. *et al.* Circuits for local and global signal integration in primary visual cortex. *J. Neurosci.* **22,** 8633–8646 (2002).
45. Koeberle, P. D. & Bahr, M. Growth and guidance cues for regenerating axons: where have they gone? *J. Neurobiol.* **59,** 162–180 (2004).
46. Fenrich, K. & Gordon, T. Canadian Association of Neuroscience review: axonal regeneration in the peripheral and central nervous systems—current issues and advances. *Can. J. Neurol. Sci.* **31,** 142–156 (2004).
47. Wandell, B. A., Chial, S. & Backus, B. Visualization and measurement of the cortical surface. *J. Cogn. Neurosci.* **12,** 739–752 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Horton and A. Wade for comments and suggestions, and C. Riedinger, R. Esaki, I. Kim, E. Lit, J. Pauls and J. Werner for technical help. We also thank L. Vaina, B. Rosen and the members of our laboratory for their advice and support. This work received support from a National Eye Institute (NEI) grant (S.M.S.), an NEI postdoctoral fellowship grant (A.S.T.), an NEI grant to B.A.W., a Howard Hughes Medical Institute Physician Postdoctoral Fellowship (S.M.S.), a National Institute of Neurological Disorders and Stroke grant (A.A.B.), a grant from the Max Planck Society and a Deutsche Forschungsgemeinschaft grant. S.M.S. is currently also affiliated with the Athinoula Martinos Center for Biomedical Imaging, Massachusetts General Hospital, CNY-Building 149, 13th Street, Charlestown, Massachusetts 02129-2000, USA.

Author Contributions S.M.S. took primary responsibility for the design and execution of all experiments, as well as for performing the analysis and preparing the manuscript. A.A.B. and M.C.S. contributed equally to this work. M.C.S. helped perform MRI experiments and analysis. A.A.B. helped with the analysis and familiarized us with the mrVISTA software. A.S.T. was intimately involved with all aspects of the work. M.A. provided technical support with the MRI experiments. A.S. provided help with the histological preparations and W.I. helped with retinal lesioning. B.A.W. and N.K.L. provided resources and acted in supervisory roles.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.M.S. (smsmirnakis@partners.org).

© 2005 Nature Publishing Group

ARTICLES

Mechanics of the kinesin step

N. J. Carter[1] & R. A. Cross[1]

Kinesin is a molecular walking machine that organizes cells by hauling packets of components directionally along microtubules. The physical mechanism that impels directional stepping is uncertain. We show here that, under very high backward loads, the intrinsic directional bias in kinesin stepping can be reversed such that the motor walks sustainedly backwards in a previously undescribed mode of ATP-dependent backward processivity. We find that both forward and backward 8-nm steps occur on the microsecond timescale and that both occur without mechanical substeps on this timescale. The data suggest an underlying mechanism in which, once ATP has bound to the microtubule-attached head, the other head undergoes a diffusional search for its next site, the outcome of which can be biased by an applied load.

Recent single-molecule work on the kinesins-1 firmly supports a model in which the motor walks along, using alternate microtubule-binding heads as holdfasts[1–3]. The step size is 8 nm, corresponding to the axial distance between microtubule heterodimer subunits, and is independent of ATP concentration and load[4,5]. There are reports of mechanical substeps[6,7]. Stepping is tightly coupled to ATP turnover: a step does not require the binding of two ATPs[8,9] and one ATP molecule is consumed on average per step[10], except possibly at high load[11].

During the long dwell intervals between steps at limiting ATP concentrations, kinesin is paused, waiting for ATP to bind. Kinesin in this waiting state is tightly attached to the microtubule by its nucleotide-free trailing (holdfast) head, whereas its ADP-containing leading (tethered) head is stably inhibited from undergoing microtubule-activated ADP release[12]. ATP binding to the holdfast head relieves this inhibition and allows the microtubule-activated release of ADP from the other head[12]. A similar scheme is thought to operate under load, such that each forward mechanical step reports the commitment of an ATP molecule to the chemical cycle[13].

Although these broad features of the mechano-chemical cycle are clear, the physical mechanism by which the motor transfers between microtubule binding sites, thereby exchanging the load from one head to the other, remains uncertain. Two different sorts of scheme have been proposed. In lever arm models, the motor head attaches tightly to the track and concomitantly or subsequently undergoes a large shift in its shape or tilt angle. The motion imparted by this conformational change can be amplified by a lever arm rigidly attached to the moving part of the motor head. Some myosins use this kind of mechanism, as evinced by a recent demonstration that the lever arm of myosin Ie could be re-engineered to harness the same conformational change to drive reverse-directed motion[14]. Recent evidence from muscle fibres confirms that the myosin II lever arm drives muscle contraction and indicates that the lever throw is dependent on load[15]. For kinesin, the neck linker model proposed by Rice *et al.*[16] is a kind of lever arm model, in which ATP-dependent docking of the flexible neck linker against the head acts like a lever arm to shift the load along the microtubule axis. A contrasting type of model emphasizes the diffusional search made by the head for its next binding site[17,18]. The motor diffuses randomly along its track, stretching a spring somewhere in its structure, and large diffusional excursions in the progress direction are selectively captured by the track-binding event. In the pure form of this kind of model, the conformational change accompanying track binding serves to stabilize binding but does not significantly stretch the spring.

The mechanism of each kinesin step could consist of elements of both processes (diffusion-to-capture plus a conformational change). Detailed information about the amplitude, time course and reversibility of stepping can help to sort out the contributions of each process. To examine stepping in more detail we used the now-classical single-bead optical-trap arrangement in which a single molecule of kinesin is attached to a spherical bead about 1 μm in diameter. The bead is optically trapped at the focus of an infrared laser[19] and the trap is steered so as to bring the kinesin within range of an immobilized microtubule. The kinesin then walks along the microtubule, tending to pull the bead out of the trap, while the trap applies an opposing force tending to restore the bead to the trap centre. Within the range of interest, the trap acts like a spring obeying Hooke's law. By accurately tracking the bead, the stepwise motions of the attached kinesin molecule can be observed. We measured in particular the effect of extreme backward loads on the pattern of stepping, and we applied a step-averaging algorithm that provides improved spatiotemporal resolution, allowing us to search for substructure within averaged 8-nm steps.

High backward loads induce processive backward stepping

At saturating ATP concentrations, when the stepping rate is not limited by the rate of ATP binding, the velocity of kinesin varies roughly linearly with force until the motor is stepping close to stall[5,20–22], at which point backward steps become more frequent (we will call steps taken towards the microtubule plus end forward steps, and steps towards the minus end backward steps). One view is that such backward steps are slippage events. Slippages and detachments certainly occur; however, it has recently been reported that the probability of single backward steps depends on the ATP concentration[17], implying that backward steps have a definite mechanism that is triggered by ATP binding, as do forward steps.

We explored the effects of substantial backward and forward loads on stepping (Fig. 1). Record 2 shows that if, during a run of processive forward steps, the load is suddenly increased to super-stall levels, kinesin can respond by stepping processively backwards until stall re-establishes. The figure shows processive runs of backward steps from 14 pN of backward force down to stall at 7 pN. At high loads, detachments are frequent, but by considering only steps that occurred within a chain of consecutive steps we were able to determine the mechanical behaviour over the entire operational range of loads.

[1]Molecular Motors Group, Marie Curie Research Institute, The Chart, Oxted, Surrey RH8 0TL, UK.

© 2005 Nature Publishing Group

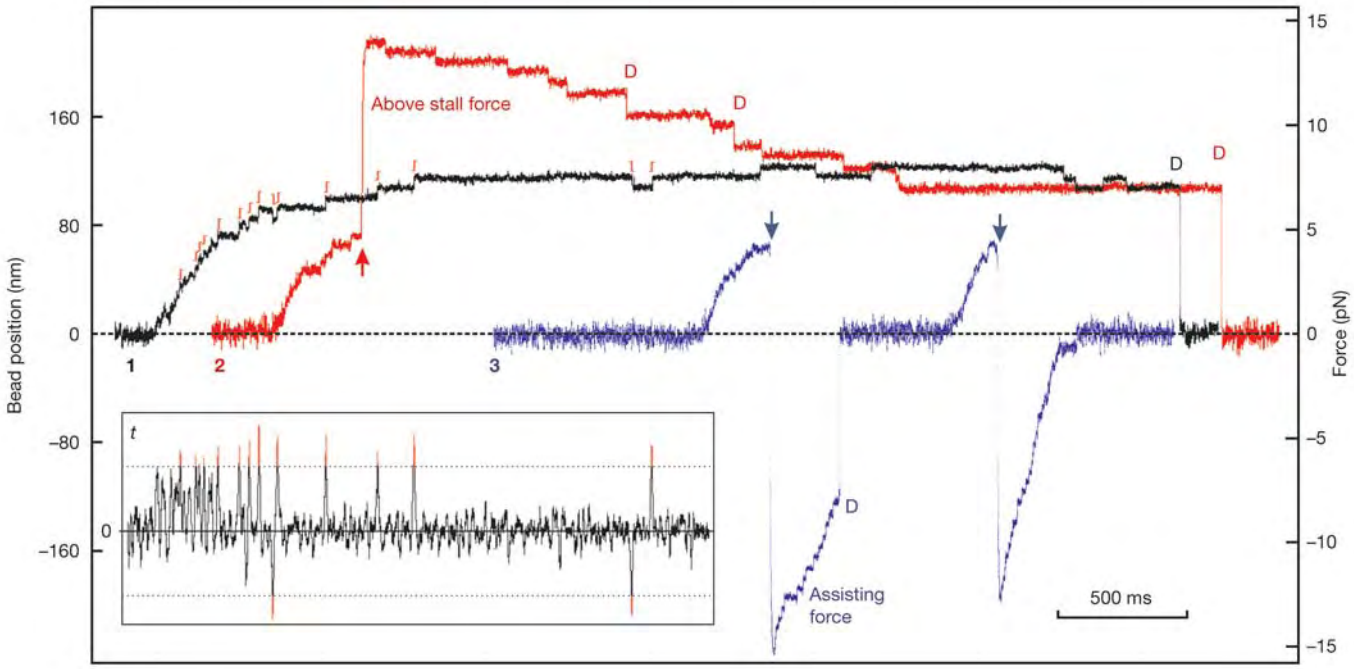


Figure 1 | Example optical trapping records. Three superimposed records showing the movement of single kinesin molecules towards stall force, over time. In record 1 (black), the trap and microtubule remain fixed throughout, and the kinesin walks with 8-nm steps away from the trap centre (dashed line) to stall force and finally detaches. In record 2 (red), on reaching 4 pN the microtubule is moved rapidly (upward arrow), pulling the kinesin to about 14 pN force. In some instances the kinesin responds with processive backward steps to stall force (7–8 pN). More commonly at forces above 10 pN, the kinesin would detach after a few or no backward steps, the bead returning to trap centre. Record 3 (blue) shows the opposite procedure. On reaching the 4-pN trigger force, the microtubule is quickly moved (downward arrows) to apply a large negative (assisting) force to the kinesin. Two successive experiments on the same molecule are shown. For automated dwell time calculations and step-averaging, a *t*-test step finder was applied to the bead position data. The inset shows the *t*-test profile for the first part of record 1. Steps are defined where the *t*-value exceeds a preset threshold value (dotted line). The located steps are shown offset just above record 1. In all records, detachment events (steps larger than 12 nm recognized by the step finder) are marked with D. Conditions: single kinesin molecules on 560-nm polystyrene beads, 1 mM ATP. The trap stiffnesses for the beads in these records were 0.064, 0.067 and 0.064 pN nm^{-1}, respectively. The force scale represents a trap stiffness of 0.065 pN nm^{-1}. Stage movements (arrows) were typically complete within 200 ms. The data shown are 1-ms boxcar filtered.

To do this we used an automatic algorithm (see Methods) to find steps within a data set of multiple records, and to determine the dwell time (the time interval between the previous step and the current step) and amplitude (the distance moved) for each step. Figure 2a shows the incidence and the amplitude of forward and backward steps over the range −15 pN (assisting load) to +15 pN (inhibitory load). The inset shows that the ratio of the number of forward steps to the number of backward steps at any particular load depends exponentially on the load. At the stall force of 7.2 pN (ref. 17) this ratio is 1. The stall force does not seem to depend on the ATP concentration, because the choice between forward and backward stepping depends only on load, not on ATP concentration.

Figure 2b plots the variation of dwell time with load for forward and backward steps. At all loads, the dwell time for both forward and backward steps decreases as the ATP concentration is increased, confirming that ATP binding is required for both forward and backward steps[17]. Under forward (assisting) loads (in the range −2 to −15 pN in Fig. 2b) only forward steps are observed, and their dwell time depends only on the ATP concentration, not on the load, indicating that a chemical step (ATP binding) is limiting. Under backward load, the dwell time for forward steps depends exponentially on the load, as expected for a particle diffusing over an activation energy barrier in accordance with Kramers theory[23–25]. The load–dwell-time curve for forward steps is exponential above about 3 pN both at 1 mM ATP and at 10 μM ATP, and the exponents are similar at high and low ATP concentrations, which is consistent with a single, common, rate-limiting mechanical step under high backward load at both high and low ATP. At any particular backward load, dwell times are exponentially distributed (Supplementary Fig. S2).

Above stall force, backward steps predominate. Their dwell times depend on the ATP concentration, confirming that backward steps, like forward steps, are triggered by ATP binding. It has been reported[17] that the mean dwell time for single backward steps is load-independent. We confirm their conclusion and extend it to include processive backward stepping under super-stall loads of up to 15 pN. Figure 2c plots force–velocity curves calculated by dividing the mean step size (forward steps positive, backward steps negative) by the mean dwell times shown in Fig. 2b.

Steps are microsecond events with no substeps

There are two reports of substeps in the literature[7,26], and several current models for the kinesin mechanism predict that the 8-nm step consists of two or more mechanical substeps, with more than one biochemical kinetic step occurring during an 8-nm physical step. To search for mechanical substeps we attached kinesin to smaller beads, to give an improved temporal response[7], and used a step-averaging algorithm to increase spatial resolution. Our algorithm automatically finds steps and then does a global fit to find the best single-exponential fit across the full data set of steps (see Methods). The origins of these fits are then used to synchronize the raw data records for the steps so that they can be ensemble averaged. We found that the resulting average step was a simple monophasic event; no mechanical substeps were detectable (Fig. 3) either before or during the major event.

To test the limits of detection in our system we generated synthetic data by adding substeps of defined amplitude and duration to real, recorded noise. The simulations (Supplementary Information) show that a bead position substep lasting more than 30 μs would be necessary to increase the averaged bead rise time detectably beyond

© 2005 Nature Publishing Group

that which we observe. Our data therefore rule out substeps of duration more than 30 μs.

Figure 3 shows that the time constants for the ensemble averaged forward and backward steps are very similar. Steps were faster if we used slightly smaller beads, which diffuse more rapidly (Fig. 3).

Model

That the forward directional bias in kinesin can be reversed by pulling backwards on the motor suggests that the tethered head makes a diffusional search for its next site, the outcome of which can be biased by an applied load.

System stiffness (trap stiffness plus bead–kinesin–microtubule link stiffnesses) in our experiments increases from about 0.06 pN nm^{-1} (trap stiffness) to more than 0.65 pN nm^{-1} during force production. All the various links in the system will contribute series compliances, but we can nevertheless use the measured stiffness at high load to set a lower limit of 0.6 pN nm^{-1} on kinesin head stiffness. This low limit for the head stiffness is consistent with a diffusional mechanism (diffusion through 8 nm followed by capture) because at zero load the Kramers time[25] for a protein of 5 nm radius to diffuse 8 nm is 7 μs. The Kramers time rises to 8 ms at a 1.7 pN nm^{-1} kinesin head stiffness, which sets an upper limit on head stiffness for a diffusional mechanism.

The absence of substeps within the 8-nm kinesin step indicates that load transfers from one head to the other in a single mechanical event. As previously discussed, we envisage that forward steps and backward steps begin from a common intermediate (the parked state; Fig. 4), in which the trailing head lacks nucleotide and acts as a holdfast to the microtubule, while the ADP-containing tethered head is unbound and unloaded[12,27]. ATP binding to the holdfast head acts as a gate that allows the tethered head to begin a diffusional search for its next binding site. Ordinarily the ensuing search pattern is biased towards the microtubule plus end. This might be achieved by having the holdfast head guiding the binding direction of its partner. The neck linker docking scheme proposed by Rice and colleagues[16] has this property, but is ruled out in its original form on several grounds[28], and particularly because neck linker docking does not have enough energy associated with it to drive stepping at high loads[29]. Nevertheless, an ATP-dependent parking and unparking of the tethered head in a forward-biased position remains in our view highly plausible, albeit in our model ATP unparks the tethered head rather than parks it. Once the tethered head locates an empty site on the microtubule, strain-dependent, microtubule-activated ADP release occurs, converting the newly attached head to strong microtubule binding and transferring the load stably to it. In the model there is a transient two-heads-attached state, but the load is always born on one head only. Backwards mechanical strain counteracts the intrinsic forward bias in the diffusional search, thus increasing the probability that the tethered head will bind behind the holdfast head. For a backward step, the newly bound tethered head is unloaded as it undergoes microtubule-activated ADP release. Consistent with this is our observation that the dwell time for backward steps at high backward loads depends on the ATP concentration but not on the load.

Comparison with other models

Competing models for the kinesin mechanism agree that a decrease in efficiency occurs close to stall, due either to futile cycles[30] or to ATP-driven backward steps[24]. In our model, tight coupling of stepping to ATP binding is retained under all conditions, but the

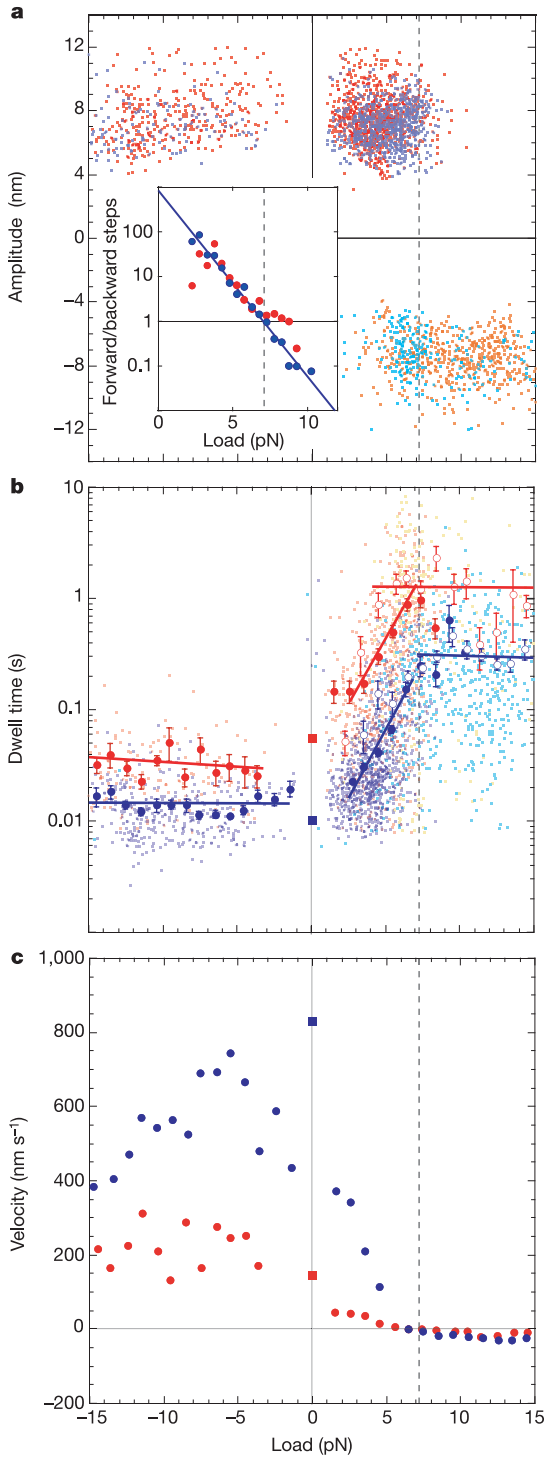


Figure 2 | Stepping behaviour. **a**, Plot of step amplitudes against load. Data for 1 mM ATP: forward steps in blue, backward steps in cyan. Data for 10 μM ATP: forward steps in red, backward steps in orange. Forward and backward steps have equal mean amplitudes. Inset, ratio of forward to backward steps plotted against the load; data for 1 mM ATP are shown in blue, and for 10 μM ATP in red. The fit is: ratio = $802e^{-0.95 \times \text{load}}$, with the load in piconewtons. **b**, Plot of dwell times against load. Data were binned with 1-pN intervals, and the mean dwell and s.e.m. were calculated for each bin. The colour key is as in **a**; filled circles are mean forward step dwells, open circles are mean backward step dwells. Mean dwell data for forward steps above 3 pN were fitted to single exponentials (solid lines). The fits are dwell = $0.0036e^{0.57\times\text{load}}$ at 1 mM ATP and dwell = $0.0256e^{0.55\times\text{load}}$ at 10 μM ATP, with the load in piconewtons. Other fits are by linear regression over the ranges shown by the solid lines. In the region 5–8 pN the dwell time distribution for backward steps is distorted slightly by the occurrence of forward steps, and vice versa (see Methods). **c**, Force–velocity curve. Velocities were calculated as mean amplitude divided by mean dwell time for each bin, taking the bins and the mean dwell times from **b**. Red data, 10 μM ATP; blue data, 1 mM ATP. The zero-load data (squares) are calculated from the bead velocities observed with the trap turned off.

© 2005 Nature Publishing Group

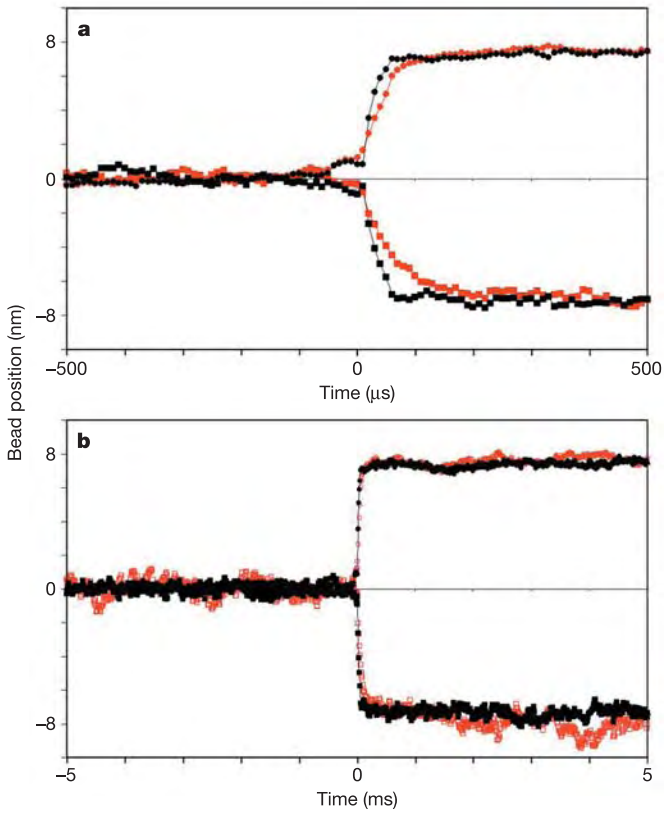


Figure 3 | Average time course of forward and backward steps. a, b, The same data sets are shown on two timescales: fine (**a**) and coarse (**b**). Step positions were determined automatically with a *t*-test step finder, followed by a least-squares exponential fit for step position refinement. The averaged forward steps (circles) and backward steps (squares) for two bead sizes are shown: the fit time constant for 500-nm beads (black) was faster than that for 800-nm beads (red). For 500-nm beads, forward steps ($n = 1{,}693$) had time constant 15.3 μs, amplitude 7.39 nm and average force 5.0 pN; for backward steps ($n = 316$) these were 19.4 μs, 7.34 nm and 6.1 pN, respectively. For 800-nm beads, forward steps ($n = 565$) had time constant 35.9 μs, amplitude 7.6 nm and force 4.7 pN; for backward steps ($n = 68$) these were 37.3 μs, 7.8 nm and 5.6 pN, respectively. All records used for the step averaging were recorded at 1 mM ATP.

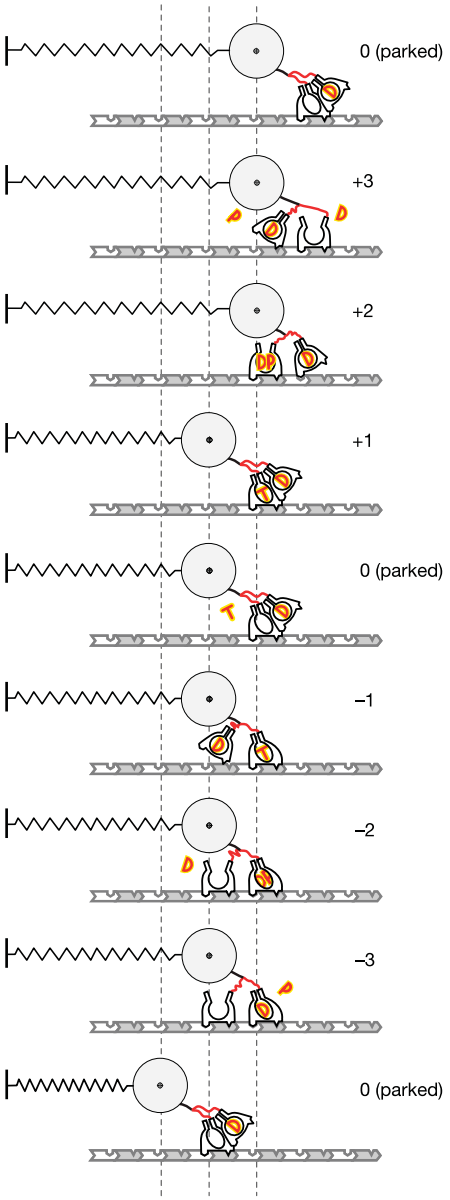


Figure 4 | Model. Before ATP binding, the motor is parked (state 0): the holdfast head remains stably bound to the microtubule and the ADP-containing tethered head cannot access its new site. Straining this state either forwards or backwards does not induce stepping. ATP (T) binding to the holdfast head sanctions ADP (D) release from the tethered head. Forward steps (+1, +2, +3, 0) occur when the tethered head binds in front of the holdfast head, backward steps (−1, −2, −3, 0) when it binds behind. The choice between forward and backward stepping depends on the applied load (the spring). In the figure, the central state 0 represents stall, in which the backward load applied by the trap (represented by the stretched spring) is such that the tethered head has an equal probability of stepping forwards or backwards. To make a forward step, the motor needs to make a diffusional excursion in the progress direction. This excursion can then be locked in by irreversible ADP release from the lead head.

probability of a forward step decreases exponentially with increasing load, whereas the probability of a backward step is constant. At stall, the probability of forward stepping is equal to that of backward stepping. Above stall, our model predicts that efficiency goes negative as backward steps begin to predominate. Note, however, that although we know that backward stepping requires ATP binding, there is at present no evidence that backward stepping is coupled to ATP turnover.

There are previous reports of substeps[6,7], and several current models predict substeps[30–32]. Our data rule out substeps that take longer than 30 μs. This indicates that the appearance of substantial substeps in previous analyses might have been an artefact. The current data gainsay models in which kinesin steps along the 4-nm tubulin monomer repeat in microtubules. Similarly, our data argue against models that predict substeps arising from rapidly equilibrating mechanical substates[30]. The data support one-stroke models[33] and specifically exclude models that forbid backward steps. Models in which backward stepping synthesizes ATP[31] are unlikely, because they predict that the frequency of backward steps should decrease at high ATP concentration, whereas we observed the opposite.

Our proposed model for processive kinesins bears striking similarities to the classical Huxley scheme for myosin[34], in that in both cases the motor head diffuses into a highly strained state, locks on to the track and then detaches at a strain-dependent rate. For myosin, this original scheme proved inadequate to account for the mechanical behaviour and was modified to include multiple bound mechanical states linked by strain-dependent conformational changes[35]. For processive kinesins, a model with biased diffusion to capture followed by a single, strain-dependent conformational change coupled to ADP release seems adequate. It will now be

© 2005 Nature Publishing Group

important to determine how the current data on the mechanism of kinesin processivity relate to the mechanisms of non-processive kinesins and of minus-end-directed kinesins.

METHODS

Sample preparation. Experiments used BRB80 plus 5 mM dithiothreitol and a glucose oxidase/catalase oxygen-scavenging system. Polystyrene beads (500, 560 and 800 nm; Polysciences) were incubated with 0.2 mg ml^{-1} casein and decreasing concentrations of full-length *Drosophila* kinesin[10]. We used bead populations in which most beads were not motile, ensuring that the vast majority of motility observed was single molecule. Pig brain microtubules were adsorbed directly on sonicator-cleaned coverslips. Each buffer contained the desired final ATP concentration (1 mM or 10 μM) to avoid any ATP dilution by mixing in the flow cell. All recordings were made at 23 °C.

Transient two-dimensional force-feedback. In some records, once the kinesin had reached a 4-pN trigger point, the motor was pulled to a large predefined positive or negative force by moving the microtubule (piezoelectric stage) in the direction of the plus or minus end. The transient software-based force-feedback uses both on-axis and off-axis bead position data, taking into account both trap stiffness and microtubule orientation and polarity. Typically, microtubule movement was complete within 200 ms, after which the feedback was switched off and the trap and stage positions were fixed.

Bead position measurement. The quadrant photodiode detector, amplifier and sampling circuitry had a combined bandwidth of 46 kHz (−3 dB position) and was free of gain-peaking. The rise time was measured at just less than 10 μs (signal equivalent to 10 nm for a 560-nm bead). The system achieves high bandwidth at the cost of increased positional noise (root-mean-square noise is equivalent to 2.1 nm of movement of the 560-nm beads, measured at 100 kHz without sample averaging). Most data were sampled at 80 kHz (per channel) and averaged down to 20 kHz for storage and analysis. A subset of the data used in the fast step averaging were sampled at 100 kHz and stored without sample averaging. With the 560-nm beads, the position detector was linear (less than 5% attenuation) over a range that extended half a bead radius on each side of the quadrant centre. For each recording, once a particular microtubule orientation and polarity had been established, the trap position was offset to provide the maximum linear range from the position detector.

Step finder algorithm: dwell times and step averaging. Kinesin steps were automatically isolated from the data by using a running *t*-test algorithm. For each data point, *W* (ms) of samples before and after that sample were compared by *t*-test. For positive force traces, *W* was 8 ms. In the faster assisting-force records, *W* was 6 ms. The resulting *t*-profile shows upward spikes for forward steps and downward spikes for backward steps (and for releases at positive forces).

Absolute *t*-values above a defined threshold were scored as steps, and the peak value was defined as time zero for their corresponding steps. An upper step-size limit of 12 nm was applied, to remove detachments and a small number of very fast double steps from consideration.

The dwell time is the time between steps. For a dwell time to be scored, both the current step and the preceding step had to fit the criteria above. We also required that the end position of the preceding step must not be more than 5 nm from the start position of the current step. This reduced the chance of erroneously scoring long dwells where steps were not resolved. We excluded steps that occurred in the range −1 to 1 pN because of the high positional noise in this region. At forces at which both forward and backward steps occurred (5–8 pN), the dwell time distribution for forward steps was distorted by the occurrence of backward steps, and vice versa.

For step averaging, further refinement of step position was performed by fitting exponentials to the steps identified by the *t*-test. In an iterative routine, the time constant that gave the minimum least-sums-of-squares fit to complete data series (typically 50–200 steps) was used to refine the step positions. For ensemble averaging, all steps were brought into register by using time zero for the fitted exponential.

Received 13 December 2004; accepted 9 March 2005.

1. Kaseda, K., Higuchi, H. & Hirose, K. Alternate fast and slow stepping of a heterodimeric kinesin molecule. *Nature Cell Biol.* **5,** 1079–1082 (2003).
2. Asbury, C. L., Fehr, A. N. & Block, S. M. Kinesin moves by an asymmetric hand-over-hand mechanism. *Science* **302,** 2130–2134 (2003).
3. Yildiz, A., Tomishige, M., Vale, R. D. & Selvin, P. R. Kinesin walks hand-over-hand. *Science* **303,** 676–678 (2004).
4. Svoboda, K., Schmidt, C. F., Schnapp, B. J. & Block, S. M. Direct observation of kinesin stepping by optical trapping interferometry. *Nature* **365,** 721–727 (1993).
5. Kojima, H., Muto, E., Higuchi, H. & Yanagida, T. Mechanics of single kinesin molecules measured by optical trapping nanometry. *Biophys. J.* **73,** 2012–2022 (1997).
6. Coppin, C. M., Finer, J. T., Spudich, J. A. & Vale, R. D. Detection of sub-8-nm movements of kinesin by high-resolution optical-trap microscopy. *Proc. Natl Acad. Sci. USA* **93,** 1913–1917 (1996).
7. Nishiyama, M., Muto, E., Inoue, Y., Yanagida, T. & Higuchi, H. Substeps within the 8-nm step of the ATPase cycle of single kinesin molecules. *Nature Cell Biol.* **3,** 425–428 (2001).
8. Hua, W., Young, E. C., Fleming, M. L. & Gelles, J. Coupling of kinesin steps to ATP hydrolysis. *Nature* **388,** 390–393 (1997).
9. Schnitzer, M. J. & Block, S. M. Kinesin hydrolyses one ATP per 8-nm step. *Nature* **388,** 386–390 (1997).
10. Coy, D. L., Wagenbach, M. & Howard, J. Kinesin takes one 8-nm step for each ATP that it hydrolyzes. *J. Biol. Chem.* **274,** 3667–3671 (1999).
11. Visscher, K., Schnitzer, M. J. & Block, S. M. Single kinesin molecules studied with a molecular force clamp. *Nature* **400,** 184–189 (1999).
12. Hackney, D. D. Evidence for alternating head catalysis by kinesin during microtubule-stimulated ATP hydrolysis. *Proc. Natl Acad. Sci. USA* **91,** 6865–6869 (1994).
13. Cross, R. A. The kinetic mechanism of kinesin. *Trends Biochem. Sci.* **29,** 301–309 (2004).
14. Tsiavaliaris, G., Fujita-Becker, S. & Manstein, D. J. Molecular engineering of a backwards-moving myosin motor. *Nature* **427,** 558–561 (2004).
15. Reconditi, M. *et al.* The myosin motor in muscle generates a smaller and slower working stroke at higher load. *Nature* **428,** 578–581 (2004).
16. Rice, S. *et al.* A structural change in the kinesin motor protein that drives motility. *Nature* **402,** 778–784 (1999).
17. Nishiyama, M., Higuchi, H. & Yanagida, T. Chemomechanical coupling of the forward and backward steps of single kinesin molecules. *Nature Cell Biol.* **4,** 790–797 (2002).
18. Okada, Y., Higuchi, H. & Hirokawa, N. Processivity of the single-headed kinesin KIF1A through biased binding to tubulin. *Nature* **424,** 574–577 (2003).
19. Block, S. M., Goldstein, L. S. & Schnapp, B. J. Bead movement by single kinesin molecules studied with optical tweezers. *Nature* **348,** 348–352 (1990).
20. Hunt, A. J., Gittes, F. & Howard, J. The force exerted by a single kinesin molecule against a viscous load. *Biophys. J.* **67,** 766–781 (1994).
21. Svoboda, K. & Block, S. M. Force and velocity measured for single kinesin molecules. *Cell* **77,** 773–784 (1994).
22. Kawaguchi, K. & Ishiwata, S. Temperature dependence of force, velocity, and processivity of single kinesin molecules. *Biochem. Biophys. Res. Commun.* **272,** 895–899 (2000).
23. Kramers, H. A. Brownian motion in a field of force and the diffusion limit of chemical reactions. *Physica* **7,** 284–304 (1940).
24. Nishiyama, M., Higuchi, H., Ishii, Y., Taniguchi, Y. & Yanagida, T. Single molecule processes on the stepwise movement of ATP-driven molecular motors. *Biosystems* **71,** 145–156 (2003).
25. Howard, J. *Mechanics of Motor Proteins and the Cytoskeleton* (Sinauer, Sunderland, Massachusetts, 2001).
26. Coppin, C. M., Pierce, D. W., Hsu, L. & Vale, R. D. The load dependence of kinesin's mechanical cycle. *Proc. Natl Acad. Sci. USA* **94,** 8539–8544 (1997).
27. Hirose, K., Lockhart, A., Cross, R. A. & Amos, L. A. Three-dimensional cryoelectron microscopy of dimeric kinesin and ncd motor domains on microtubules. *Proc. Natl Acad. Sci. USA* **93,** 9539–9544 (1996).
28. Schief, W. R. & Howard, J. Conformational changes during kinesin motility. *Curr. Opin. Cell Biol.* **13,** 19–28 (2001).
29. Rice, S. *et al.* Thermodynamic properties of the Kinesin neck-region docking to the catalytic core. *Biophys. J.* **84,** 1844–1854 (2003).
30. Schnitzer, M. J., Visscher, K. & Block, S. M. Force production by single kinesin motors. *Nature Cell Biol.* **2,** 718–723 (2000).
31. Fisher, M. E. & Kolomeisky, A. B. Simple mechanochemistry describes the dynamics of kinesin molecules. *Proc. Natl Acad. Sci. USA* **98,** 7748–7753 (2001).
32. Thomas, N., Imafuku, Y., Kamiya, T. & Tawada, K. Kinesin: a molecular motor with a spring in its step. *Proc. R. Soc. Lond. B* **269,** 2363–2371 (2002).
33. Block, S. M., Asbury, C. L., Shaevitz, J. W. & Lang, M. J. Probing the kinesin reaction cycle with a 2D optical force clamp. *Proc. Natl Acad. Sci. USA* **100,** 2351–2356 (2003).
34. Huxley, A. F. Muscle structure and theories of contraction. *Prog. Biophys. Biophys. Chem.* **7,** 257–318 (1957).
35. Huxley, A. F. & Simmons, R. M. Proposed mechanism of force generation in striated muscle. *Nature* **233,** 533–538 (1971).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Howard for the gift of the *Drososphila* kinesin, the reviewers of this manuscript for careful and constructive criticism, and Marie Curie Cancer Care for unswerving support.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.A.C. (r.cross@mcri.ac.uk).

© 2005 Nature Publishing Group

LETTERS

Detection and imaging of atmospheric radio flashes from cosmic ray air showers

H. Falcke[1,4,9], W. D. Apel[6], A. F. Badea[6], L. Bähren[4], K. Bekk[6], A. Bercuci[3], M. Bertaina[11], P. L. Biermann[1], J. Blümer[5,6], H. Bozdog[6], I. M. Brancus[3], S. Buitink[9], M. Brüggemann[10], P. Buchholz[10], H. Butcher[4], A. Chiavassa[11], K. Daumiller[6], A. G. de Bruyn[4], C. M. de Vos[4], F. Di Pierro[11], P. Doll[6], R. Engel[6], H. Gemmeke[7], P. L. Ghia[12], R. Glasstetter[13], C. Grupen[10], A. Haungs[6], D. Heck[6], J. R. Hörandel[5], A. Horneffer[1], T. Huege[1], K.-H. Kampert[13], G. W. Kant[4], U. Klein[2], Y. Kolotaev[10], Y. Koopman[3], O. Krömer[7], J. Kuijpers[9], S. Lafebre[9], G. Maier[6]†, H. J. Mathes[6], H. J. Mayer[6], J. Milke[6], B. Mitrica[3], C. Morello[12], G. Navarra[11], S. Nehls[6], A. Nigl[9], R. Obenland[6], J. Oehlschläger[6], S. Ostapchenko[6], S. Over[10], H. J. Pepping[4], M. Petcu[3], J. Petrovic[9], S. Plewnia[6], H. Rebel[6], A. Risse[8], M. Roth[5], H. Schieler[6], G. Schoonderbeek[4], O. Sima[3], M. Stümpert[5], G. Toma[3], G. C. Trinchero[12], H. Ulrich[6], S. Valchierotti[11], J. van Buren[6], W. van Cappellen[4], W. Walkowiak[10], A. Weindl[6], S. Wijnholds[4], J. Wochele[6], J. Zabierowski[8], J. A. Zensus[1] & D. Zimmermann[10]

The nature of ultrahigh-energy cosmic rays (UHECRs) at energies >10^{20} eV remains a mystery[1]. They are likely to be of extragalactic origin, but should be absorbed within ~50 Mpc through interactions with the cosmic microwave background. As there are no sufficiently powerful accelerators within this distance from the Galaxy, explanations for UHECRs range from unusual astrophysical sources to exotic string physics[2]. Also unclear is whether UHECRs consist of protons, heavy nuclei, neutrinos or γ-rays. To resolve these questions, larger detectors with higher duty cycles and which combine multiple detection techniques[3] are needed. Radio emission from UHECRs, on the other hand, is unaffected by attenuation, has a high duty cycle, gives calorimetric measurements and provides high directional accuracy. Here we report the detection of radio flashes from cosmic-ray air showers using low-cost digital radio receivers. We show that the radiation can be understood in terms of the geosynchrotron effect[4–8]. Our results show that it should be possible to determine the nature and composition of UHECRs with combined radio and particle detectors, and to detect the ultrahigh-energy neutrinos expected from flavour mixing[9,10].

When UHECRs interact with particles in the Earth's atmosphere, they produce a shower of elementary particles propagating towards the ground with almost the speed of light. The first suggestion that these air showers also could produce radio emission was made[11] on the basis of a charge-excess mechanism, which is very strong for showers developing in solid media[12,13]. In a couple of experimental activities in the 1960s and 1970s, coincidences between radio pulses and cosmic ray events were indeed reported[14,15]. Owing to the limitations of electronics in those days, the measurements were cumbersome and did not lead to useful relations between radio emission and air shower parameters. As a consequence, the method was not pursued for a long time and the historic results came into question. However, the mechanism for the radio emission of air showers was recently revisited and proposed to be coherent geosynchrotron emission[4]: Secondary electrons and positrons produced in the particle cascade rush with velocities close to the speed of light through the Earth's magnetic field and are deflected. As in synchrotron radiation, this produces dipole radiation that is relativistically beamed into the forward direction. The shower front emitting the radiation has a thickness that is comparable to (or less than) a wavelength for radio emission below ~100 MHz. Hence the emission is expected to be coherent to a large extent, which greatly amplifies the signal.

To see whether radio emission from cosmic rays is indeed detectable and useful in a modern cosmic ray experiment, we have built the LOPES (LOFAR Prototype Station) experiment. LOPES is a phased array of dipole antennas with digital electronics developed to test aspects of the LOFAR (Low-Frequency Array) concept. Compared to historical experiments, it provides an order of magnitude increase in bandwidth and time resolution, effective digital filtering methods, and for the first time true interferometric imaging capabilities. The radio array is co-located with the KASCADE[16] (Karlsruhe Shower Core and Array Detector) experiment that is now part of KASCADE-Grande at the research centre in Karlsruhe, Germany (see Supplementary Fig. 1). KASCADE provides coincidence triggers for LOPES and well-calibrated information about air shower properties. Experimental procedure and data reduction have been described by Horneffer *et al.*[17], and we give here only a brief summary in the Methods section. A related experiment is currently under way at the Nançay radio observatory[18].

Using LOPES, we have detected the radio emission from cosmic ray air showers at 43–73 MHz on a regular basis with unsurpassed spatial and temporal resolution (see Supplementary Fig. 2). After digital filtering of radio interference we still have almost the full bandwidth of $\Delta\nu = 33$ MHz available, giving us a time resolution of $\Delta t \approx 1/\Delta\nu = 30$ ns compared to ~1 μs resolution in historic experiments. Using radio interferometric techniques we can also image the radio flash for the first time. An example is shown in Fig. 1, where the air shower is the brightest radio point source on the sky for some tens of nanoseconds (see also Supplementary Video 1). The nominal

[1]Max-Planck-Institut für Radioastronomie, 53121 Bonn, Germany. [2]Radioastronomisches Institut der Universität Bonn, 53121 Bonn, Germany. [3]National Institute of Physics and Nuclear Engineering, 7690 Bucharest, Romania. [4]ASTRON, 7990 AA Dwingeloo, The Netherlands. [5]Institut für Experimentelle Kernphysik, Universität Karlsruhe, 76021 Karlsruhe, Germany. [6]Institut für Kernphysik, Forschungszentrum Karlsruhe, 76021 Karlsruhe, Germany. [7]Institut für Prozessdatenverarbeitung und Elektronik, FZK, 76021 Karlsruhe, Germany. [8]Soltan Institute for Nuclear Studies, 90950 Lodz, Poland. [9]Department of Astrophysics, Radboud University, 6525 ED Nijmegen, The Netherlands. [10]Fachbereich Physik, Universität Siegen, 57072 Siegen, Germany. [11]Dipartimento di Fisica Generale dell'Università, 10125 Torino, Italy. [12]Istituto di Fisica dello Spazio Interplanetario, INAF, 10133 Torino, Italy. [13]Fachbereich C - Physik, Universität Wuppertal, 42097 Wuppertal, Germany. †Present address: University of Leeds, Leeds LS2 9JT, UK.

© 2005 Nature Publishing Group

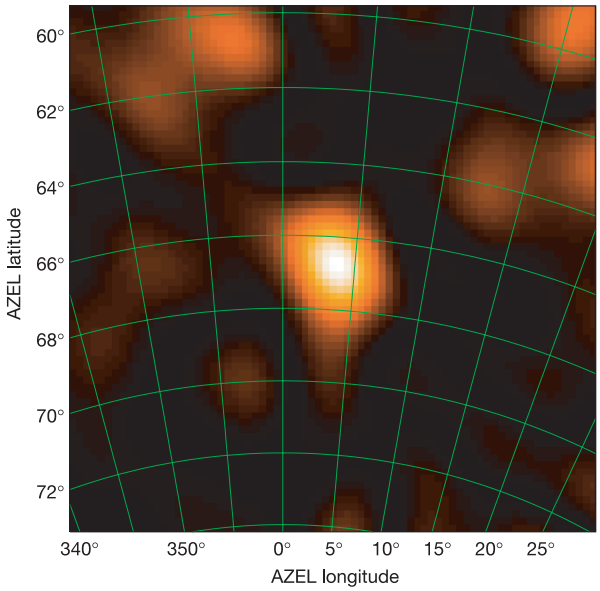


Figure 1 | Radio map of an air shower. For each pixel in the map, we formed a beam in this direction integrated over 12.5 ns and show the resulting electric field intensity. Longitude and latitude give the azimuth and elevation (AZEL) direction (north is to the top, east to the right). The map is focused towards a distance of 2,000 m (fixed curvature radius for each pixel). The cosmic ray event is seen as a bright blob of 2.4° × 1.8° size. Most of the noise in this map is due to interferometer sidelobes caused by the sparse radio array. No image deconvolution has been performed.

resolution in our maps is ~2° in azimuth and elevation towards the zenith. Within these limits the emission appears point-like. To put this in perspective, we note that previous radio experiments used fixed analogue beams with a width of ~20° and no possibilities for imaging. Current detectors for UHECRs (for example, AUGER) are limited to about ~1° accuracy. Positional accuracies for LOPES can be a fraction of a degree for bright sources. This will improve further with interferometer baselines longer than used here.

To make an initial and reliable statistical assessment of the radio properties of air showers, we have investigated a rather restrictive set of events with relatively high signal-to-noise ratio and simple selection criteria. The criteria are purely based on shower parameters reconstructed from KASCADE. Using events from the first half year of operation, starting January 2004, we selected all events with a shower core within 70 m of the centre of LOPES, a zenith angle <45°, and a reconstructed 'truncated muon number' of $N_\mu > 10^{5.6} \approx 4 \times 10^5$. The truncated muon number is the reconstructed number of muons within 40–200 m of the shower core. For KASCADE, this quantity is a good tracer of primary particle energy[19], $E_p \propto N_\mu^{0.9}$. The selection corresponds approximately to $E_p > 10^{17}$ eV. This is the upper end of the energies that KASCADE was built for. Using these selection criteria ('cuts') leaves us with 15 events and a 100% detection efficiency of the radio signal. This avoids any bias due to non-detections. The rather restrictive cut on the shower core location allows us to ignore radial dependencies. Also, the antenna gain reduces significantly at zenith angles >45°. Even though neither KASCADE nor LOPES are optimized for large zenith angles, we have also checked for highly inclined events. Selecting all events with a much lower truncated muon number of $N_\mu > 10^5$ and zenith angle >50° we still detect >50% of all events—in many cases with very high field strengths. However, given the current uncertainties of KASCADE in shower parameters for inclined showers, we ignore those events for our analysis below.

The strength of the detected radio pulses in our sample is some μV per m per MHz at present, but we still lack an accurate absolute gain calibration. The position of the radio flashes are coincident with the direction of the shower axis derived from KASCADE data within the errors. The average offset is (0.8 ± 0.4)° as determined from our radio maps.

We find the strongest correlations between the absolute value of the electric field strength height of the pulse, ε and N_μ, and between ε and the geomagnetic angle, α_B. The latter is defined as the angle between the shower axis and the geomagnetic field. No obvious correlation is found with the zenith angle. Theory[4] predicts that owing to coherence the electric field strength should scale linearly with the number of particles in the shower, which—for KASCADE—is approximately proportional to the number of muons. Hence, to first order we can separate the two effects by dividing the electric field of the radio pulse by the truncated muon number. We find that $\varepsilon_\mu = \varepsilon/N_\mu \propto (1 - \cos\alpha_B)$, where the lowest geomagnetic angle in the sample was $\alpha_B = 8°$ (see Supplementary Fig. 3). It is also possible to use a $\varepsilon_\mu \propto \sin\alpha_B$ behaviour, which for our data gives only marginally worse results. Simulations[6,7] indicate that this strong dependence on α_B is largely a polarization effect, as LOPES antennas currently measure only a single polarization in the east–west direction.

We can use this dependence to correct for the geomagnetic angle. Figure 2a shows the measured radio signal ε plotted against muon number, while Fig. 2b shows the same plot where we use a normalized field strength height, ε_α, corrected for the $(1 - \cos\alpha_B)$ dependence.

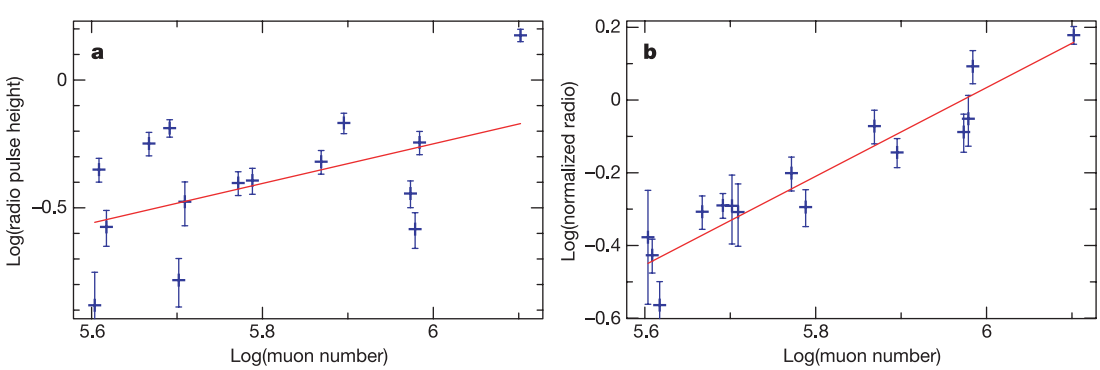


Figure 2 | Radio emission as a function of muon number. **a**, The logarithm of the radio pulse height, ε, versus the logarithm of muon number, which has not been corrected for the geomagnetic angle dependence. **b**, As **a** but now the geomagnetic angle dependence is corrected for (ε_α). The correlation improves significantly. Solid lines indicate power-law fits. Errors were calculated from the noise in the time series before the pulse plus a nominal 5% error on gain stability. Both errors were added in quadrature. The radio pulse height units are arbitrary. Errors are one-sigma standard deviations.

© 2005 Nature Publishing Group

The correlation, albeit with only few events, clearly improves and shows a remarkably low scatter of ±16%. For an error-weighted nonlinear fit we find that $\varepsilon_\alpha \propto N_\mu^{1.2\pm0.1}$, which, within the errors, is consistent with a linear relation, $\varepsilon_\alpha \propto E_p$.

The close association between radio flashes and cosmic ray air showers in direction and time shows that the radio emission is directly associated with the shower itself. Our selection of events based purely on shower parameters has yielded a 100% detection efficiency, suggesting that radio emission from extensive air showers is a common and reliable tracer of cosmic rays. As predicted, the electric field of the radio emission is coherent. The good correlation with the geomagnetic angle demonstrates that the emission is caused by the interaction of the shower with the Earth's magnetic field. All results that we have found so far match the basic predictions for the geosynchrotron effect.

The most encouraging result is the very tight correlation of the radio emission with the primary particle energy. The essentially linear increase of the electric field strength with energy is a consequence of coherence, and hence the received radio power will increase quadratically with primary particle energy, making this technique particularly suitable for cosmic rays at much higher energies than studied here. For covering large detector areas—to increase count rates at the high-energy end of the UHECR spectrum—radio antennas are particularly suitable, owing to their ease of deployment and low cost.

Radio detection of cosmic rays with digital phased arrays has a number of features that make it superior (or complementary) to current techniques for UHECR detection in a number of areas. For example, the good directional accuracy reachable with radio interferometers will significantly improve clustering studies in search for UHECR point sources if used in a large ground array like LOFAR.

Another very promising area involves composition studies. The number of electrons integrated over the entire shower, revealing the primary particle energy, is difficult to determine with particle detectors: only a small fraction of these electrons reach the ground, due to their short absorption length. Muons, on the other hand, reach the ground largely unharmed, and their number is higher for showers produced by an iron nucleus compared to proton-induced, or even γ-induced, showers. Hence, a combined radio- and muon-detector array could determine the spectrum and composition of UHECRs. Compared to hybrid detector arrays, where optical fluorescence techniques with ~10% duty cycle are used to obtain calorimetric measurements, the duty cycle for radio hybrid studies would be almost ten times higher. Hence radio arrays could be the way to study UHECR composition at the very highest energies, where much larger event rates are needed than are currently available.

A particularly intriguing application concerns highly inclined showers, which were expected to be very easily detected with radio antennas[5,6,20]. The predicted detectability is supported by our high detection rate of showers at large zenith angles, despite a lower threshold on muon number. Highly inclined showers travel through several times the vertical air mass, and very few shower particles reach a detector on the ground. The exceptions are neutrinos, which can travel large distances without interactions and generate inclined showers at any distance to the ground[9]. Showers induced by electron neutrinos will have a relatively low hadronic component relative to the leptonic component. Therefore, for inclined showers the ratio between the radio and the muon signal on the ground will be a tell-tale signal of electron neutrinos.

Additionally, it has been suggested that Earth-skimming tau neutrinos, produced through flavour mixing, will also lead to highly inclined air showers when the secondary tau decays after some 50 km travel length at 10^{18} eV (ref. 10). No hadronic or photon showers are expected at such low inclinations. Hence, radio antennas with sensitivity towards the horizon would also be ideal tau neutrino detectors and provide clues about flavour mixing. Although the first detection of ultrahigh-energy neutrino events would be extremely exciting in itself, the calorimetric and far-looking nature of radio detection would even allow a relatively reliable energy determination of electron and tau neutrino showers—something that had been considered extremely difficult in the past with surface detector arrays[10].

METHODS

Description of the array and standard data processing. The current set-up of LOPES consists of 10 'inverted-V' antennas distributed over the KASCADE array[16], which consists of 252 detector stations on a uniform grid with 13 m spacing, and which is electronically organized in 16 independent clusters. Each antenna is centred between four KASCADE huts in the northwestern part of the array. The antenna set-up has a maximum baseline of 125 m.

The analogue radio signal received by the antenna is filtered to select a band from 43 to 76 MHz, giving a bandwidth of $\Delta\nu = 33$ MHz, and is digitized with a 12-bit, 80-MHz analogue-to-digital converter (ADC), that is, using the second Nyquist zone (40–80 MHz) of the ADC. The data are continuously written into a cyclic ring buffer with a buffering time of 6.7 s (1 Gbyte). About 0.8 ms of data are read out whenever KASCADE produces a 'large-event trigger', which means that 10 out of 16 KASCADE clusters have produced an internal trigger. The resulting trigger rate is about two per minute, giving a total data volume of about 3.5 Gbyte d^{-1}. The dead-time during read-out is ~0.6 s. On average the trigger is delayed with respect to the shower by 1.8 μs, depending on the shower geometry relative to the array.

We relate the radio signal to shower properties using parameters provided by the standard KASCADE data processing from the particle detectors[16]. These parameters are: location of the shower core, shower direction, energy deposited in the particle detectors, total number of electrons at ground level, and the reconstructed truncated muon number.

The basic processing of the data consists of several steps: correction of instrumental delays for each antenna using a TV transmitter with known position, digital filtering of narrow-band interference, frequency and elevation-dependent gain correction for each antenna, flagging of antennas with unusually high noise, and correction of trigger delays based on shower direction. Delay corrections are done by applying a phase gradient to the Fourier transform of the time series data. Digital filtering is achieved by setting the amplitudes of narrow-band transmitters in the Fourier domain to an average value of the surrounding channels[17]. In the next step, a digital beam is formed in the direction of the shower axis by correcting for the delay of the light-travel time, and summing up the digitized and calibrated E-field time series data from each antenna ('adding beam'). Alternatively, one can also add up data from each unique combination of antennas where the E-field time series data of the antennas in each pair have been multiplied ('cross-correlation beam') before summing. To obtain the received power, the adding beam needs to be squared. Conversion back to 'historically' used units[15] of $\mu V\,m^{-1}\,MHz^{-1}$—an integrated absolute pulse field strength height ε—is done by multiplying with $1/\Delta\nu$ and taking the square root of the data.

We fit a gaussian profile to the largest peak in a narrow time window around −1.8 μs from the trigger signal where the shower is expected to arrive. Finally, we check for the curvature of the wave front by correcting antenna delays assuming a spherical wave front with curvature radius R_{cur}, and iterating R_{cur} until the radio peak is maximal.

The average full-width at half-maximum of the radio pulses in time was 49 ± 10 ns. This is mostly due to broadening by the finite filter bandwidth, which we have not yet attempted to deconvolve. The average curvature radius of the wave front was $R_{cur} = 5{,}000 \pm 3{,}000$ m, the latter being a statistical scatter. The curvature is rather poorly determined, as our longest baselines are an order of magnitude smaller than the average R_{cur}. Curvature has only a significant effect for events with high signal-to-noise ratio. For weak events, the change in amplitude for different R_{cur} is less than the noise in the pulse height. An analysis using a fixed value of $R_{cur} = 2{,}500$ m for each event was found to change the final result in this Letter only marginally.

In addition to the standard data reduction described above, we produced for each of the selected events a sky-map of the radio emission. The mapping routine calculates the radio intensity by adding the electric fields of the dipoles for a four-dimensional data grid as a function of azimuth, elevation, distance and time. We used a spatial resolution of 0.2° on the plane of the sky, distance slices separated by 250 m, and a time resolution of 12.5 ns. Within this four-dimensional data cube we located the intensity maximum to obtain the radio properties of the shower, and fitted a gaussian to the time profile. The resolution of the interferometer beam (~2°) becomes asymmetric and degrades in elevation for sources towards the horizon, as we have a planar array.

© 2005 Nature Publishing Group

Received 14 February; accepted 4 April 2005.

1. Nagano, M. & Watson, A. A. Observations and implications of the ultrahigh-energy cosmic rays. *Rev. Mod. Phys.* **72,** 689–732 (2000).
2. Sigl, G. The enigma of the highest energy particles of nature. *Ann. Phys.* **303,** 117–141 (2003).
3. Haungs, A., Rebel, H. & Roth, M. Energy spectrum and mass composition of high-energy cosmic rays. *Rep. Prog. Phys.* **66,** 1145–1206 (2003).
4. Falcke, H. & Gorham, P. W. Detecting radio emission from cosmic ray air showers and neutrinos with a digital radio telescope. *Astropart. Phys.* **19,** 477–494 (2003).
5. Huege, T. & Falcke, H. Radio emission from cosmic ray air showers. Monte Carlo simulations. *Astron. Astrophys.* **430,** 779–798 (2005).
6. Huege, T. & Falcke, H. Radio emission from cosmic ray air showers: simulation results and parametrization. *Astropart. Phys.* (submitted); preprint at ⟨http://arxiv.org/astro-ph/0501580⟩ (2005).
7. Huege, T. & Falcke, H. Radio emission from cosmic ray air showers. Coherent geosynchrotron radiation. *Astron. Astrophys.* **412,** 19–34 (2003).
8. Suprun, D. A., Gorham, P. W. & Rosner, J. L. Synchrotron radiation at radio frequencies from cosmic ray air showers. *Astropart. Phys.* **20,** 157–168 (2003).
9. Capelle, K. S., Cronin, J. W., Parente, G. & Zas, E. On the detection of ultra high energy neutrinos with the Auger observatory. *Astropart. Phys.* **8,** 321–328 (1998).
10. Bertou, X., Billoir, P., Deligny, O., Lachaud, C. & Letessier-Selvon, A. Tau neutrinos in the Auger Observatory: a new window to UHECR sources. *Astropart. Phys.* **17,** 183–193 (2002).
11. Askaryan, G. A. Excess negative charge of an electron-photon shower and its coherent radio emission. *Sov. Phys. JETP* **14,** 441–443 (1962).
12. Saltzberg, D. *et al.* Observation of the Askaryan effect: Coherent microwave Cherenkov emission from charge asymmetry in high-energy particle cascades. *Phys. Rev. Lett.* **86,** 2802–2805 (2001).
13. Zas, E., Halzen, F. & Stanev, T. Electromagnetic pulses from high-energy showers: Implications for neutrino detection. *Phys. Rev. D* **45,** 362–376 (1992).
14. Jelley, J. V. *et al.* Radio pulses from extensive cosmic-ray air showers. *Nature* **205,** 327–328 (1965).
15. Allan, H. R. Radio emission from extensive air showers. *Prog. Element. Part. Cosmic Ray Phys.* **10,** 169–302 (1971).
16. Antoni, T. *et al.* The cosmic-ray experiment KASCADE. *Nucl. Instrum. Methods A* **513,** 490–510 (2003).
17. Horneffer, A. *et al.* LOPES: Detecting radio emission from cosmic ray air showers. *Proc. SPIE* **5500,** 129–138 (2004).
18. Ravel, O. *et al.* Radio detection of cosmic ray air showers by the CODALEMA experiment. *Nucl. Instrum. Methods Phys. A* **518,** 213–215 (2004).
19. Antoni, T. *et al.* A non-parametric approach to infer the energy spectrum and the mass composition of cosmic rays. *Astropart. Phys.* **16,** 245–263 (2002).
20. Gousset, T., Ravel, O. & Roy, C. Are vertical cosmic rays the most suitable to radio detection? *Astropart. Phys.* **22,** 103–107 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements A.F.B. is on leave of absence from the National Institute of Physics and Nuclear Engineering, Bucharest; T.H. is now at the Institut für Kernphysik, Forschungszentrum Karlsruhe, Karlsruhe; S. Ostapchenko is on leave of absence from Moscow State University, Moscow. LOPES was supported by the German Federal Ministry of Education and Research (Verbundforschung Astroteilchenphysik). This work is part of the research programme of the Stichting voor Fundamenteel Onderzoek der Materie (FOM), which is financially supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). The KASCADE experiment is supported by the German Federal Ministry of Education and Research. The Polish group is supported by KBN; the Romanian group acknowledge support from the Romanian Ministry of Education and Research.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.F. (falcke@astron.nl).

© 2005 Nature Publishing Group

LETTERS

CO self-shielding as the origin of oxygen isotope anomalies in the early solar nebula

J. R. Lyons[1,2] & E. D. Young[1,2]

The abundances of oxygen isotopes in the most refractory mineral phases (calcium-aluminium-rich inclusions, CAIs) in meteorites[1] have hitherto defied explanation. Most processes fractionate isotopes by nuclear mass; that is, ^{18}O is twice as fractionated as ^{17}O, relative to ^{16}O. In CAIs ^{17}O and ^{18}O are nearly equally fractionated, implying a fundamentally different mechanism. The CAI data were originally interpreted as evidence for supernova input of pure ^{16}O into the solar nebula[1], but the lack of a similar isotope trend in other elements argues against this explanation[2]. A symmetry-dependent fractionation mechanism[3,4] may have occurred in the inner solar nebula[5], but experimental evidence is lacking. Isotope-selective photodissociation of CO in the innermost solar nebula[6] might explain the CAI data, but the high temperatures in this region would have rapidly erased the signature[7]. Here we report time-dependent calculations of CO photodissociation in the cooler surface region of a turbulent nebula. If the surface were irradiated by a far-ultraviolet flux $\sim 10^3$ times that of the local interstellar medium (for example, owing to an O or B star within $\sim$1 pc of the protosun), then substantial fractionation of the oxygen isotopes was possible on a timescale of $\sim 10^5$ years. We predict that similarly irradiated protoplanetary disks will have H_2O enriched in ^{17}O and ^{18}O by several tens of per cent relative to CO.

Self-shielding of CO in the solar nebula is a process of isotope-selective photodissociation that occurs at far-ultraviolet (FUV) wavelengths from 91.2 nm to 110 nm. Because CO first goes to a bound excited state before dissociating, the absorption spectrum of CO consists of many narrow lines. The wavelength of each line is determined by the specific vibrational and rotational levels involved, and is shifted when the mass of the molecule is changed as a result of an isotopic substitution. The absorption spectra of the various CO isotopologues do not overlap significantly, particularly at the low temperatures of molecular clouds and the outer regions of the solar nebula. During photodissociation the most abundant isotopologue ($C^{16}O$) saturates, which reduces its rate of dissociation relative to the less abundant $C^{18}O$ and $C^{17}O$. This produces a zone of enrichment of ^{18}O and ^{17}O, and corresponding depletion of $C^{18}O$ and $C^{17}O$.

Very large fractionations ($\sim 10^4$‰) are observed[8,9] and predicted[10] in diffuse and translucent molecular clouds, and it has recently been suggested[11] that H_2O produced by CO dissociation in the parent molecular cloud can explain oxygen isotope abundances in CAIs. Because large fractionations are not observed in cloud cores[9], this mechanism implies that the solar nebula derived from cloud envelope material. However, star formation results from core collapse in dense molecular clouds, and it is unclear whether large fractionations in more tenuous clouds are inherited by protostellar nebulae.

Surface regions of the nebula are similar in temperature and pressure to dense molecular clouds, and are therefore a likely site for formation and preservation of oxygen isotope heterogeneity produced during self-shielding[12]. To investigate the influence of CO self-shielding on oxygen isotope ratios in the early Solar System, we used a one-dimensional photochemical model to compute the time-dependent oxygen isotope profiles in a two-dimensional, axisymmetric nebula. The total gas number density in the disk was specified with a standard analytical disk model[13]. For a nebula of bulk composition similar to the solar composition[14], the initial volume fractions are $f_{He} = 0.16$, $f_{CO} = 2 \times 10^{-4}$, $f_{H_2O} = 2 \times 10^{-4}$ and $f_{MgFeSiO_4} = 2 \times 10^{-5}$, assuming that all C is in CO, all Si is in $MgFeSiO_4$, and all remaining O is in H_2O.

Disk photochemistry in the model is initiated by CO dissociation

$$C^xO + h\nu \rightarrow C + {}^x O \qquad (1)$$

where $x = 16$, 17 and 18, and by H_2 and H photoionization at wavelengths less than 91.2 nm. Disk chemistry was restricted to H-, C- and O-containing chemical species, and the reaction rate coefficients used are a subset of the UMIST kinetics database[15]. By using previously published disk models[13,16,17], we developed a reduced set of 96 species and 375 reactions to model the disk chemistry. Gas–grain reactions are also included, using published desorption energies[16,17]. Although our model is less sophisticated in its treatment of chemistry and radiative transfer than some recent models[18,19], it is, to our knowledge, the first model to include both vertical transport and oxygen isotopes.

To follow the time evolution and vertical distribution of O-containing species in the disk resulting from CO photodissociation high above the midplane, we solved the one-dimensional continuity equation for each species as a function of height above the midplane (see Supplementary Information). Vertical motion is characterized by a vertical diffusion coefficient, D_v, which is assumed to be comparable to the turbulent viscosity, $\nu_t = \alpha c H$, where c is the sound speed, H is the vertical scale height in the nebular gas, and $\alpha < 1$ is a parameter used in disk models to describe the strength of turbulent mixing[20]. Theoretical models of disk instabilities[20] suggest that $\alpha \approx 10^{-3}$ to 10^{-1}. All quantities depend on r, the radial location in the disk, and so a value of r must be specified. We have chosen $r = 30$ AU (midplane temperature of 51 K) as representative of cold regions of the disk where water ice formation is expected.

To quantify the effects of self-shielding, we derived fits to numerically determined shielding functions developed for molecular clouds in the interstellar medium[10,21] (see Supplementary Information). The photodissociation rate of each CO isotopologue is proportional to the product εJ_{ISM}, where $J_{ISM} = 2.0 \times 10^{-10}\ s^{-1}$ is the rate constant for CO photolysis due to the local interstellar medium (LISM) FUV field[10], and ε is an FUV flux enhancement factor. Dust opacity is identical for all isotopologues, and was parameterized in the same manner as is done for the interstellar medium. We consider a range of $\varepsilon = 1$ to 10^5, consistent with an FUV enhancement due to proximity to a massive O star in a star-forming region[22].

Integration of the system of continuity equations yields the time evolution of species abundances in the model. Following CO

[1]Institute of Geophysics and Planetary Physics, [2]Department of Earth and Space Sciences, University of California, Los Angeles, California 90095, USA.

© 2005 Nature Publishing Group

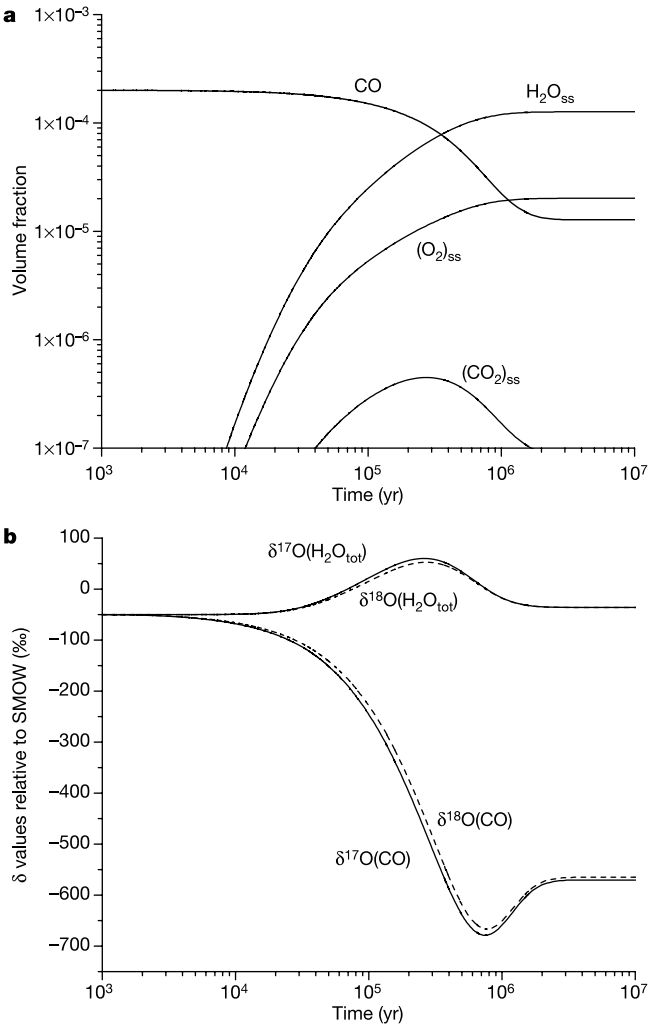


Figure 1 | Model results for the time evolution of molecular abundances and isotope ratios at the nebula midplane at a heliocentric distance of 30 AU. The temperature at the midplane is 51 K, and 0.1-μm dust particles are uniformly distributed in the gas. The turbulent viscosity parameter $\alpha = 10^{-2}$, consistent with previously published estimates[20,26]. The FUV flux enhancement factor is $\varepsilon = 500$ relative to the LISM flux, which implies a dust temperature of 110 K at the nebula surface. **a**, Volume fractions of several species relative to total nebular gas (the subscript 'ss' refers to O derived from CO self-shielding). C is ionized and forms a suite of hydrocarbon ions and molecules (not shown) as is predicted in other disk models[17]. **b**, $\delta^{17}O$ and $\delta^{18}O$ of CO and total nebular H_2O (H_2O_{tot}) relative to standard mean ocean water (SMOW). H_2O_{tot} consists of H_2O_{ss} plus H_2O from the parent molecular cloud, H_2O_{cloud}. The oxygen isotope δ-value for an unknown 'u' is computed as

$$\delta^x O_{CO_{initial}}(u) = 10^3 \left(\frac{(^xO/^{16}O)_u}{(^xO/^{16}O)_{CO_{initial}}} - 1 \right)$$

for all oxygen-containing molecules in the nebula, for $x = 17$ and 18. Initial CO in the nebula is assumed to have oxygen isotope ratios of $^{16}O/^{18}O = 500$ and $^{16}O/^{17}O = 2{,}600$. The δ-values are then converted from a 'CO initial' reference to a SMOW reference (see Supplementary Information), assuming that $\delta^x O_{SMOW}(CO_{initial}) = -50‰$, the measured value of the isotopically lightest CAIs. The oxygen isotope composition of total nebular H_2O is then given by the weighted sum of $\delta^x O_{SMOW}(H_2O_{cloud})$ and $\delta^x O_{SMOW}(H_2O_{ss})$, where the volume fraction of H_2O from the parent cloud is $f_{H_2O_{cloud}} = 2 \times 10^{-4}$ and the isotope composition of the H_2O in the cloud is $\delta^x O_{SMOW}(H_2O_{cloud}) = -50‰$.

self-shielding and dissociation, H_2O is formed by reactions of H and O on grain surfaces. The formation timescale for H_2O is $\sim 10^5$ years at the midplane for $\alpha = 10^{-2}$ and $\varepsilon = 500$ (Fig. 1a). Total water in the nebula, H_2O_{tot}, consists of H_2O from the parent molecular cloud and the H_2O produced from O liberated during CO self-shielding, H_2O_{ss}. For comparison with meteorite oxygen isotope data, the model oxygen isotope ratios ($^{18}O/^{16}O$ and $^{17}O/^{16}O$) are expressed as δ-values ($\delta^{17}O_{SMOW}$ and $\delta^{18}O_{SMOW}$) relative to the isotope ratios in ocean water (see Fig. 1 legend). A difference of several hundred per mil is predicted between the $\delta^{17}O_{SMOW}$ and $\delta^{18}O_{SMOW}$ values of nebular H_2O and CO at times $>10^4$ years (Fig. 1b). The non-mass-dependent isotope signature, $\Delta^{17}O_{SMOW} = \delta^{17}O_{SMOW} - 0.52\,\delta^{18}O_{SMOW}$, is a useful quantity because it is unaffected by mass-dependent fractionation processes. An FUV flux enhancement of $\varepsilon \approx 10^2$–10^3 yields $\Delta^{17}O_{SMOW}$ of H_2O_{tot} in the range inferred from meteorites on timescales much less than the lifetime of gas in the nebula, which is $\sim 10^6$–10^7 years (Fig. 2).

Figure 3 shows the trajectory of $\delta^{17}O_{SMOW}$ and $\delta^{18}O_{SMOW}$ values of nebular water at the midplane for $\varepsilon = 500$. The slope of the trajectory is ~5–10% too high compared to the slope-0.94 line[2] and slope-1.0 line[23] measured for CAIs. A slope >1 is expected for self-shielding in pure CO because of the greater abundance, and therefore greater self-shielding, of $C^{18}O$ compared to $C^{17}O$. Because CO dissociation occurs in an H_2-rich environment, the wavelength dependence of H_2 absorption must be included in CO self-shielding calculations. About 60% of the dissociation of $C^{18}O$ in molecular clouds occurs[10]

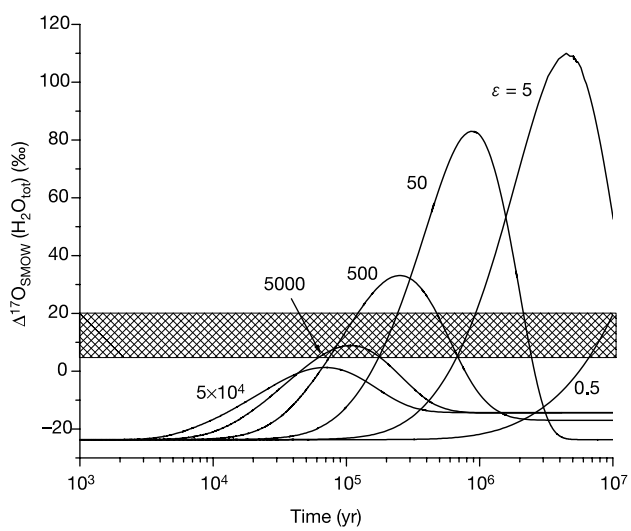


Figure 2 | The time evolution of the non-mass-dependent oxygen isotope component of total nebular H_2O for a range of FUV flux enhancement factors. Other parameters are as given in Fig. 1. The non-mass-dependent isotope signature is defined as $\Delta^{17}O_{SMOW} = \delta^{17}O_{SMOW} - 0.52 \times \delta^{18}O_{SMOW}$. The cross-hatched area indicates the range of minimum $\Delta^{17}O_{SMOW}$ inferred for total nebular H_2O from analyses of carbonaceous chondrites[27–29]; higher $\Delta^{17}O_{SMOW}$ values of total nebula water are allowed by the meteorite data. For $\varepsilon = 5$, 50 and 500, $\Delta^{17}O_{SMOW}(H_2O_{tot})$ values at the midplane are within or above the cross-hatched region. Lower FUV fluxes do not photolyse enough CO within 1 Myr to produce sufficient fractionation at the midplane. For a given α, the maximum $\Delta^{17}O_{SMOW}(H_2O_{tot})$ decreases as FUV flux increases, suggesting that it may be possible to place an upper limit on the FUV incident upon the disk; for this case, the upper limit would be $\varepsilon \approx 10^4$. The isotopic composition of nebular H_2O is also dependent on α. Results for $\alpha = 10^{-3}$ (not shown) are similar to $\alpha = 10^{-2}$, but timescales are ~3 times longer. For $\alpha = 10^{-4}$, weak mixing implies an insufficient turnover of CO to maintain efficient self-shielding, yielding $\Delta^{17}O_{SMOW}(H_2O)$ well below the values inferred from meteorites. Thus, the value of $\Delta^{17}O_{SMOW}$ of nebular H_2O predicted by the model is sensitive to a key disk parameter, α, and to a key nebular environment parameter, the FUV flux incident on the disk.

© 2005 Nature Publishing Group

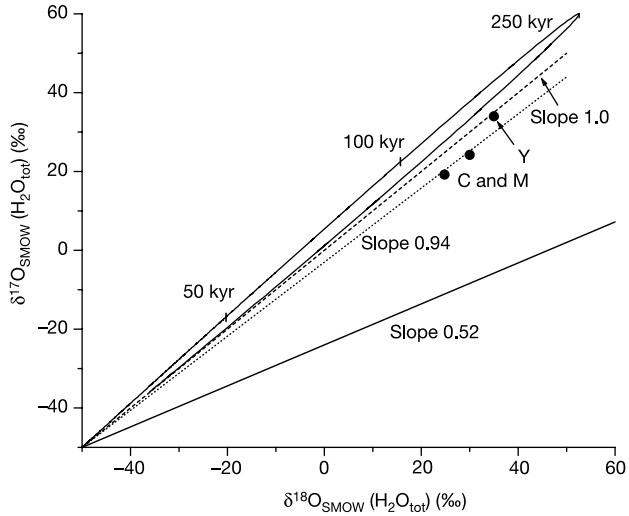


Figure 3 | Three-isotope plot ($\delta^{17}O_{SMOW}$ versus $\delta^{18}O_{SMOW}$) of total nebular H_2O at the midplane, with time labelled along the trajectory. Model parameters are as given in Fig. 1. 'Slope' is defined by the ratio $\delta^{17}O/\delta^{18}O$. The carbonaceous chondrite anhydrous mineral (CCAM) line[1,2] (slope 0.94), which describes bulk CAIs, is shown for reference. A line of slope 1.00, measured in a particularly unaltered CAI of the Allende meteorite[23], and a mass-dependent fractionation line (slope 0.52), are also shown. The minimum isotopic composition of initial nebular H_2O as inferred from analyses of carbonaceous chondrites is indicated by 'C and M' (ref. 27) and 'Y' (ref. 29). The slope of the model trajectory is ~10% too high compared to the CCAM line, and ~5% higher than the slope-1.0 line. A similar slope was estimated[30] for self-shielding in pure O_2. A slope >1 is expected for the model trajectory, because although absorption by H_2 is accounted for (see Supplementary Information), the wavelength dependence of the overlap of H_2 absorption on CO isotopologue lines has not been included.

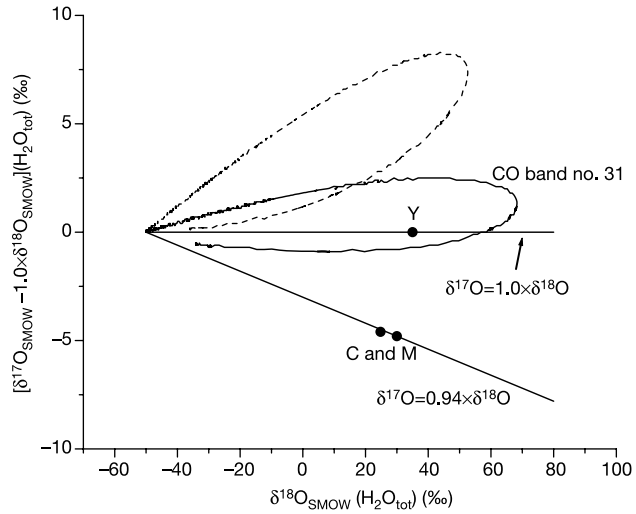


Figure 4 | Three-isotope plot of total nebular H_2O in the model, including an estimate of wavelength-dependent absorption by H_2. The ordinate is $\delta^{17}O - 1.0 \times \delta^{18}O$, rather than $\delta^{18}O$, to emphasize differences in slope. The unlabelled curve (dashed line) is the same as the trajectory shown in Fig. 3. The solid curve includes differential shielding (that is, wavelength-dependent shielding) by H_2 on CO band number 31. CO band 31 accounts for ~60% of $C^{18}O$ photodissociation in translucent molecular clouds[10]. Shielding functions that account for the differential shielding by H_2 on the isotopologues of CO band 31 were computed from H_2 synthetic spectra (see Supplementary Information). Differential shielding by H_2 brings the model δ-values for total nebular water within the range inferred from meteorites[27,29]. The differential shielding analysis shown here applies to CO self-shielding in both the surface region of the disk and in the parent molecular cloud.

in a particular band of CO (band no. 31). Inclusion of H_2 absorption on band 31 (see Supplementary Information) brings the model slope into the range inferred for nebular water, and provides strong support for the self-shielding mechanism (Fig. 4).

Transfer of the non-mass-dependent signature in nebular water to the rocky component of the nebula requires H_2O to have been concentrated relative to CO in the disk midplane[11]. Without concentration of H_2O at the midplane, equilibration of CO and H_2O as material migrated inward to ~1–2 AU would have returned both CO and H_2O to the bulk isotopic values of the parent cloud. For the model presented here, concentration of ice-coated dust with $\Delta^{17}O > 0$ at the midplane implies grain growth timescales of $>10^4$ years. Inward migration of ice-rich metre-sized objects may have then delivered ^{17}O and ^{18}O-rich water to the snowline[24]. However, layered accretion[25] is likely to have been important in the ionized, self-shielding region of the disk, with the result that $\Delta^{17}O$-rich water was deposited along the top of the disk dead zone (a predicted zone of weak mixing). Quantitative evaluation of oxygen isotopes in a layered accretion disk will require a two-dimensional chemical transport model.

We conclude that if CO self-shielding occurred primarily in the solar nebula, a strong FUV source was present (~10^3 times LISM) and disk vertical mixing was vigorous ($\alpha \approx 10^{-2}$). An FUV flux of this magnitude is expected from an O or early B star within a distance of ~1 pc of the protosun, and implies solar birth in a cluster of ~200 stars[22], a very plausible birth scenario for our Solar System. The requirement of vigorous vertical mixing is consistent with theoretical predictions of the dynamics of disk instabilities[20,25]. Our model makes two testable predictions. First, as pointed out by Clayton[6], solar oxygen isotopes will be similar to the isotopically lightest CAIs ($\delta^{17}O_{SMOW} \approx \delta^{18}O_{SMOW} \approx -50$‰), a prediction that will be tested by analysis of solar wind samples returned by the GENESIS spacecraft. Second, H_2O in similarly irradiated protoplanetary disks will be enriched in ^{17}O and ^{18}O by ~30–100% relative to disk CO, a difference that will be measurable by the next generation of infrared and submillimetre telescopes (for example, SOFIA, Herschel and ALMA).

Received 27 December 2004; accepted 10 March 2005.

1. Clayton, R. N., Grossman, L. & Mayeda, T. K. Component of primitive nuclear composition in carbonaceous meteorites. *Science* **182**, 485–488 (1973).
2. Clayton, R. N. Oxygen isotopes in meteorites. *Annu. Rev. Earth Planet. Sci.* **21**, 115–149 (1993).
3. Thiemens, M. H. & Heidenreich, H. E. Mass independent fractionation of oxygen: a novel isotope effect and its possible cosmochemical implications. *Science* **219**, 1073–1075 (1983).
4. Gao, Y. Q. & Marcus, R. A. Strange and unconventional isotope effects in ozone formation. *Science* **293**, 259–263 (2001).
5. Marcus, R. A. Mass-independent isotope effect in the earliest processed solids in the solar system: A possible chemical mechanism. *J. Chem. Phys.* **121**, 8201–8211 (2004).
6. Clayton, R. N. Self-shielding in the solar nebula. *Nature* **415**, 860–861 (2002).
7. Lyons, J. R. & Young, E. D. Towards an evaluation of self-shielding at the X-point as the origin of the oxygen isotope anomaly in CAIs. *Lunar Planet. Sci. Conf.* **XXXIV**, abstr. 1981 (2003).
8. Bally, J. & Langer, W. D. Isotope-selective photodissociation of carbon monoxide. *Astrophys. J.* **255**, 143–148 (1982).
9. Federman, S. R. *et al.* Further evidence for chemical fractionation from ultraviolet observations of carbon monoxide. *Astrophys. J.* **591**, 986–999 (2003).
10. van Dishoeck, E. F. & Black, J. H. The photodissociation and chemistry of interstellar CO. *Astrophys. J.* **334**, 771–802 (1988).
11. Yurimoto, H. & Kuramoto, K. Molecular cloud origin for the oxygen isotope heterogeneity in the solar system. *Science* **305**, 1763–1766 (2004).
12. Lyons, J. R. & Young, E. D. Evolution of oxygen isotopes in the solar nebula. *Lunar Planet. Sci. Conf.* **XXXV**, abstr. 1970 (2004).
13. Aikawa, Y. & Herbst, E. Two-dimensional distributions and column densities of gaseous molecules in protoplanetary disks. *Astron. Astrophys.* **371**, 1107–1117 (2001).

© 2005 Nature Publishing Group

14. Allende Prieto, C., Lambert, D. L. & Asplund, M. A reappraisal of the solar photospheric C/O ratio. *Astrophys. J.* **573,** L137–L140 (2002).
15. Le Teuff, Y. H., Millar, T. J. & Markwick, A. J. The UMIST database for astrochemistry 1999. *Astron. Astrophys. Suppl. Ser.* **146,** 157–168 (2000).
16. Hasegawa, T. I., Herbst, E. & Leung, C. M. Models of gas-grain chemistry in dense interstellar clouds with complex organic molecules. *Astrophys. J.* **82,** 167–195 (1992).
17. Willacy, K., Klahr, H. H., Millar, T. J. & Henning, Th. Gas and grain chemistry in a protoplanetary disk. *Astron. Astrophys.* **338,** 995–1005 (1998).
18. van Zadelhoff, G. J., Aikawa, Y., Hogerheijde, M. R. & van Dishoeck, E. F. Axi-symmetric models of ultraviolet radiative transfer with applications to circumstellar disk chemistry. *Astron. Astrophys.* **397,** 789–802 (2003).
19. Markwick, A. J., Ilgner, M., Millar, T. J. & Henning, Th. Molecular distributions in the inner regions of protostellar disks. *Astron. Astrophys.* **385,** 632–646 (2002).
20. Stone, J. M., Gammie, C. F., Balbus, S. A. & Hawley, J. F. in *Protostars and Planets IV* (eds Manning, V., Boss, A. P. & Russell, S. S.) 589–611 (Univ. Arizona Press, Tucson, 2000).
21. Lee, H. H., Herbst, E., Pineau des Forêts, G., Roueff, E. & Le Bourlot, J. Photodissociation of H_2 and CO and time dependent chemistry in inhomogeneous interstellar clouds. *Astron. Astrophys.* **311,** 690–707 (1996).
22. Adams, F. C., Hollenbach, D., Laughlin, G. & Gorti, U. Photoevaporation of circumstellar disks due to external far-ultraviolet radiation in stellar aggregates. *Astrophys. J.* **611,** 360–379 (2004).
23. Young, E. D. & Russell, S. S. Oxygen reservoirs in the early solar nebula inferred from an Allende CAI. *Science* **282,** 452–455 (1998).
24. Cuzzi, J. N. & Zahnle, K. J. Material enhancement in protoplanetary nebulae by particle drift through evaporation fronts. *Astrophys. J.* **614,** 490–496 (2004).
25. Gammie, C. F. Layered accretion in T Tauri disks. *Astrophys. J.* **457,** 355–362 (1996).
26. Cuzzi, J. N., Dobrovolskis, A. R. & Hogan, R. C. in *Chondrules and the Protoplanetary Disk* (eds Hewins, R. H., Jones, R. H. & Scott, E. R. D.) 35–43 (Cambridge Univ. Press, Cambridge, 1996).
27. Clayton, R. N. & Mayeda, T. K. The oxygen isotope record in Murchison and other carbonaceous chondrites. *Earth Planet. Sci. Lett.* **67,** 151–161 (1984).
28. Choi, B. G., Krot, A. N., McKeegan, K. D. & Wasson, J. T. Extreme oxygen-isotope compositions in magnetite from unequilibrated ordinary chondrites. *Nature* **392,** 577–579 (1998).
29. Young, E. D. The hydrology of carbonaceous chondrite parent bodies and the evolution of planet progenitors. *Phil. Trans. R. Soc. Lond. A* **359,** 2095–2110 (2001).
30. Navon, O. & Wasserburg, G. J. Self-shielding in O_2: a possible explanation for oxygen isotope anomalies in meteorites? *Earth Planet. Sci. Lett.* **73,** 1–16 (1985).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements J.R.L. thanks K. McKeegan, J. Cuzzi and A. Boss for discussions, and P. Plavchan for assistance with IDL. J.R.L. acknowledges funding from the NASA Origins Program and from the UCLA Center for Astrobiology. E.D.Y. acknowledges support from the UCLA Center for Astrobiology.

Author Contributions J.R.L. conceived and carried out the calculations presented here, using a modified version of a photochemical code provided by J. Kasting. The paper was written by J.R.L. and E.D.Y.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.R.L. (jrl@ess.ucla.edu).

© 2005 Nature Publishing Group

LETTERS

An optical lattice clock

Masao Takamoto[1], Feng-Lei Hong[3], Ryoichi Higashi[1] & Hidetoshi Katori[1,2]

The precision measurement of time and frequency is a prerequisite not only for fundamental science but also for technologies that support broadband communication networks and navigation with global positioning systems (GPS). The SI second is currently realized by the microwave transition of Cs atoms with a fractional uncertainty of 10^{-15} (ref. 1). Thanks to the optical frequency comb technique[2,3], which established a coherent link between optical and radio frequencies, optical clocks[4] have attracted increasing interest as regards future atomic clocks with superior precision. To date, single trapped ions[4–6] and ultracold neutral atoms in free fall[7,8] have shown record high performance that is approaching that of the best Cs fountain clocks[1]. Here we report a different approach, in which atoms trapped in an optical lattice serve as quantum references. The 'optical lattice clock'[9,10] demonstrates a linewidth one order of magnitude narrower than that observed for neutral-atom optical clocks[7,8,11], and its stability is better than that of single-ion clocks[4,5]. The transition frequency for the Sr lattice clock is 429,228,004,229,952(15) Hz, as determined by an optical frequency comb referenced to the SI second.

Accurate atomic clocks rely on the observation of a narrow atomic resonance $\Delta\nu$ at a transition frequency ν_0 that is insensitive to external perturbations to the highest possible degree. An indicator of clock performance is the fractional instability, which is minimized by repeatedly measuring the high-Q ($Q = \nu_0/\Delta\nu$) transition. The fractional instability is given by the Allan deviation[12]

$$\sigma_y(\tau) \approx \Delta\nu/(\nu_0\sqrt{N\tau}) \qquad (1)$$

where N is the total number of oscillators (atoms or ions) measured in unit time and τ is the total measurement time. To improve the stability of the current Cs clock, the formula suggests the use of a transition with a higher frequency ν_0 (that is, use of higher frequency than that of microwaves), and this has led to active research towards optical clocks[4–8,11]. The use of well-designed atom traps may further increase the stability, since the extended coherent interaction time Δt reduces the Fourier limit linewidth as $\Delta\nu \approx 1/\Delta t$, thus allowing a higher effective Q-factor.

A single quantum absorber held in a space smaller than the transition wavelength[13] (that is, in the Lamb-Dicke regime) would offer the ultimate system for ultra-precise spectroscopy, as such a trap significantly decreases both atomic interactions and the Doppler shifts that critically affect clock accuracy[7,8]. This system is actually embodied by an ion in a Paul trap, where the ion is trapped at the zero of a quadrupole electric field and is not perturbed by the trapping field[14], and this has provided the finest optical spectrum yet obtained[4]. However, strong Coulomb interactions between ions prevent the use of more than a single ion in a trap, which severely limits clock stability because of small N in equation (1).

In this regard, neutral atoms with much weaker interactions[7,8,11] are suitable in terms of increasing the number of particles, and therefore the stability. A spatial interference pattern of lasers can produce periodic trapping potentials for ultracold neutral atoms, called an optical lattice[15], as illustrated in Fig. 1a. This lattice potential can confine atoms in a submicrometre region, and its periodicity allows the production of billions of micro-traps in a volume of 1 mm^3. These features are indeed attractive for fine spectroscopy with enhanced stability.

In general, such a lattice-trapping field significantly modifies the internal states of atoms by so-called light shifts, and so the system was not seriously considered for atomic clocks until the demonstration of the light shift cancellation technique[16,17]. The transition frequency ν

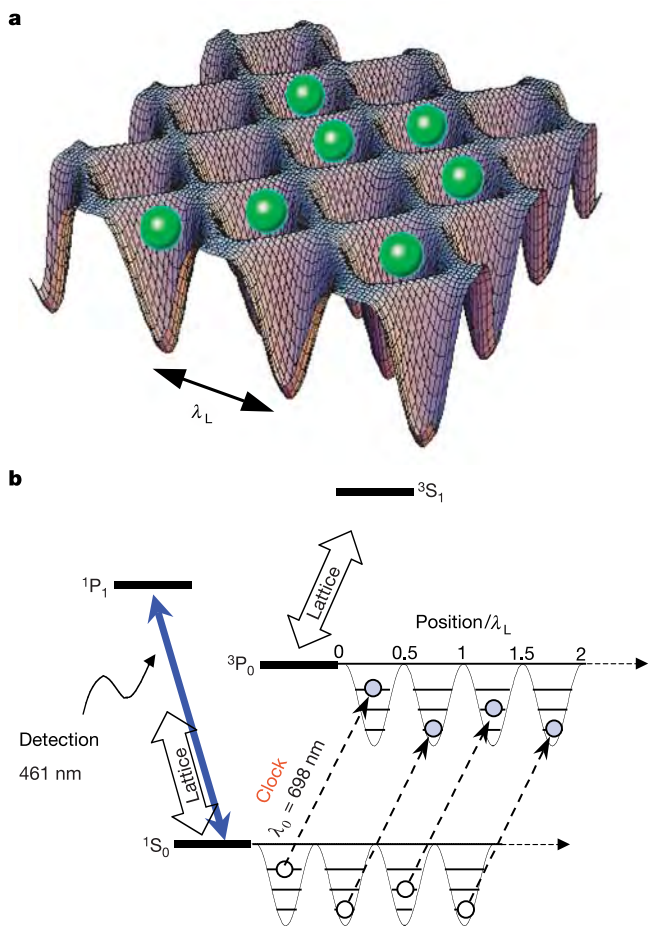


Figure 1 | Optical lattice clock. a, The spatial interference pattern of lasers creates a lattice potential that confines atoms in a region much smaller than the optical wavelength, λ_L. **b**, Energy levels for Sr. The $^1S_0$ and $^3P_0$ states are coupled to the upper respective spin states by an off-resonant laser to produce an optical lattice with equal energy shifts in the clock transition at $\lambda_0 = 698$ nm. Atoms are excited on the $|^1S_0\rangle\otimes|n\rangle \rightarrow |^3P_0\rangle\otimes|n\rangle$ electronic–vibrational transitions, where n denotes the vibrational states of atoms in the lattice potential. The clock transition is then monitored on the $^1S_0 - {}^1P_1$ cyclic transition with nearly unit quantum efficiency.

[1]Engineering Research Institute, The University of Tokyo and [2]PRESTO, Japan Science and Technology Agency, Bunkyo-ku, Tokyo 113-8656, Japan. [3]National Metrology Institute of Japan/National Institute of Advanced Industrial Science and Technology (NMIJ/AIST), Tsukuba, Ibaraki 305-8563, Japan.

© 2005 Nature Publishing Group

of atoms exposed to a lattice laser with intensity I is given by the sum of the unperturbed transition frequency ν_0 and the light shift ν_{ac}:

$$\nu(\lambda_L, \mathbf{e}) = \nu_0 + \nu_{ac} = \nu_0 - \frac{\Delta\alpha(\lambda_L, \mathbf{e})}{2\varepsilon_0 ch} I + O(I^2) \qquad (2)$$

Here ε_0, c and h are the permittivity of free space, the speed of light and the Planck constant, respectively, and $\Delta\alpha(\lambda_L, \mathbf{e}) = \alpha_u(\lambda_L, \mathbf{e}) - \alpha_l(\lambda_L, \mathbf{e})$ is the difference between the a.c. polarizabilities of the upper and lower states (α_u and α_l, respectively), which depends both on the lattice laser wavelength λ_L and its polarization vector $\mathbf{e}$. If we adjust these polarizabilities to satisfy $\Delta\alpha(\lambda_L, \mathbf{e}) = 0$, the observed atomic transition frequency ν will be equal to ν_0 independent of the lattice laser intensity I, as long as the higher-order corrections $O(I^2)$ are negligible. In the search for a clock transition in which the light shift cancellation condition $\Delta\alpha(\lambda_L, \mathbf{e}) = 0$ can be given mainly by the laser wavelength λ_L, it is preferable to use the $J = 0$ state, which exhibits a scalar light shift and does not depend on the light polarization. These 'tricks' applied here rely on the fact that the frequency or wavelength is the most accurately measurable physical quantity, which we used to avoid measuring laser intensity or polarization.

In this experiment, we adopted the $5s^2\ {}^1S_0\ (F = 9/2) \rightarrow 5s5p\ {}^3P_0\ (F = 9/2)$ forbidden transition of ^{87}Sr with a nuclear spin of 9/2 as the clock transition[9] (Fig. 1b), in which hyperfine mixing provides a finite lifetime of $\Gamma_0^{-1} = 150$ s. Our calculations[10] have shown that the ${}^1S_0$ and ${}^3P_0$ states provide an equal light shift for a lattice laser tuned to $\lambda_L \approx 800$ nm. Around this 'magic' wavelength, the light shift can be controlled with 1 mHz precision by defining the lattice laser frequency only within 1 MHz. Other systematic effects, such as the vector and tensor light shift arising from the hyperfine splitting, are summarized in the right column of Table 1 and in ref. 10. In these calculations, we assumed the typical lattice laser intensity of $I = 10\ \mathrm{kW\,cm^{-2}}$.

The experimental configuration for the lattice clock is shown in Fig. 2a (ref. 18). ^{87}Sr atoms were laser-cooled and trapped on the ${}^1S_0 - {}^3P_1$ transition. Roughly 10^4 atoms with a temperature of ~2 μK were loaded into a 20-μK-deep one-dimensional optical lattice that was formed by the standing wave of a Ti:sapphire laser (spatially filtered by an optical fibre) and provided atoms with subwavelength confinement along the axial direction. A clock laser operating at $\lambda_0 = 698$ nm with a linewidth of ~10 Hz was sent into the same fibre to set its wave vector parallel to the axis of the one-dimensional lattice potential so as to satisfy the Lamb-Dicke condition. Both lasers were passed through a polarizer with an extinction ratio of 10^{-5} to guarantee linear polarization. We note that for this π-polarized lattice laser, the frequency spread due to the tensor as well as the vector light shift among Zeeman sublevels m is less than 1 mHz (ref. 10; see Table 1). The first-order Zeeman shift for $\Delta m = 0$ transitions is $m \times 106\ \mathrm{Hz\,G^{-1}}$. The residual magnetic field of 22.5 mG and random population distribution among m substates could be responsible for the Zeeman shift or broadening up to 10 Hz in this experiment. The reduction of the magnetic field down to the μG level would be feasible by applying magnetic field shielding.

We irradiated the clock laser for 10–40 ms to excite atoms trapped in the optical lattice to the ${}^3P_0$ state. The excitation of the clock transition was then monitored by laser-induced fluorescence on the ${}^1S_0 - {}^1P_1$ cyclic transition. The 'magic' wavelength for the lattice laser was determined as follows. We measured the lattice laser intensity (I)-dependent light shift ν_{ac} at several wavelengths λ_L near 813.5 nm (ref. 18), as shown in Fig. 3a. The slope of the fitted lines, which gives the differential light shift $\delta\nu_{ac}(\lambda_L, I)/\delta I$ and is proportional to the polarizability difference $\Delta\alpha(\lambda_L)$ (see equation (2)), is plotted as a function of the lattice laser wavelength λ_L in Fig. 3b. The 'magic' wavelength, where $\Delta\alpha(\lambda_L) = 0$ holds, is therefore determined as 813.420(7) nm. We note that this uncertainty introduces the scalar-light-shift uncertainty of 4 Hz (see Table 1).

The inset in Fig. 4 shows the spectrum of the clock transition at the 'magic' wavelength; each open circle was measured with an

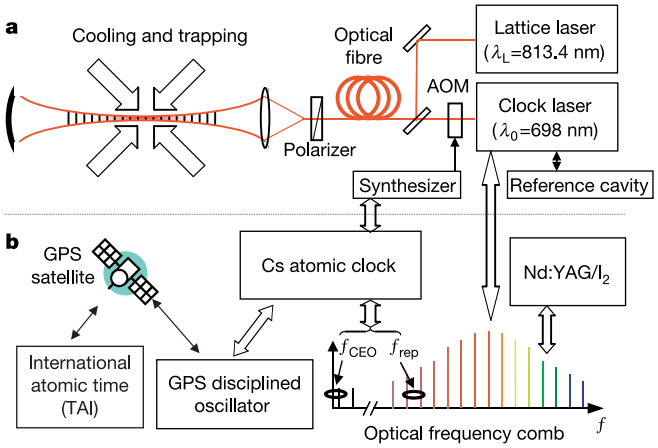


Figure 2 | Experimental configuration. a, Ultracold ^{87}Sr atoms are loaded into a one-dimensional optical lattice produced by the standing wave of a Ti:sapphire laser tuned to the 'magic' wavelength. The atoms are confined at the anti-nodes of the standing wave, which give subwavelength confinement along the axial direction. The atoms interact with the clock laser propagating along this axis, and the Lamb-Dicke condition is satisfied. AOM, acousto-optic modulator. **b**, The clock laser frequency is linked to a commercial Cs atomic clock and an iodine-stabilized Nd:YAG laser by using the optical frequency comb, where f_{CEO} and f_{rep} denote the carrier-envelope offset frequency and the repetition rate of the laser pulse, respectively. The offset frequency of the Cs clock is referenced to international atomic time through GPS time.

Table 1 | Systematic corrections and uncertainties for the Sr optical lattice clock

Effect	Correction (Hz)	Uncertainty (Hz)	
		Achieved	Attainable
First order Doppler*	0	3×10^{-2}	$<10^{-3}$
Second order Doppler	0	2×10^{-6}	$<2 \times 10^{-6}$
Recoil shift	0		
First order Zeeman	0	10	10^{-3}
Collision shift	0.6	2.4	10^{-5}†
Blackbody shift‡	2.4	0.1	3×10^{-3}
Probe laser light shift	0.1	0.01	10^{-3}
Scalar light shift	−3.8§	4	10^{-3}
Vector light shift‖	0	10^{-3}	10^{-3}
Tensor light shift	0	10^{-3}	10^{-3}
Fourth order light shift¶	0	10^{-3}	10^{-3}
Cs clock offset	−45	3	
Frequency measurement	0	9	
Systematic total	−45.7		
Total uncertainty, $\delta\nu$		15	4×10^{-3}

* The first order Doppler shift may occur owing to the relative motion between the clock laser and the lattice potential, which is often introduced by the vibration of the optical fibre or air turbulence where the clock laser travels. Taking the clock laser linewidth (~10 Hz) to be this Doppler width, the frequency uncertainty may decrease down to 30 mHz after averaging 10^5 measurements. By applying fibre phase-noise cancellation, by referring a reflection from a mirror surface that forms the standing wave for the lattice potential, this Doppler shift could be actively stabilized to less than 1 mHz (ref. 28).

† The resonant dipole–dipole interaction between atoms separately confined in the three-dimensional (3D) lattice potential with an interatomic distance of $\lambda_L/2 \approx 400$ nm is assumed[10].

‡ The blackbody shift scales as $\nu_B \approx 3 \times 10^{-10} \times T^4$ Hz (where T is in units of K), which would introduce 3 mHz uncertainty for the temperature inhomogeneity of $\Delta T = 0.1$ K at $T = 293$ K.

§ The measurements in Fig. 4 were carried out with the lattice laser wavelength at 813.4270 nm, which was finally found to be 0.007 nm longer than the actual 'magic' wavelength. The Stark shift introduced by this wavelength difference is corrected.

‖ In realizing 3D optical lattices, one can apply a phase stable optical lattice where the 3D laser intersection is formed by a single linearly polarized standing-wave with folded mirrors[29]. In this configuration, since the amplitude and the phase of the trapping light fields are correlated in each lattice site, the high polarization stability is expected.

¶ This is estimated without including the resonant contribution[10].

© 2005 Nature Publishing Group

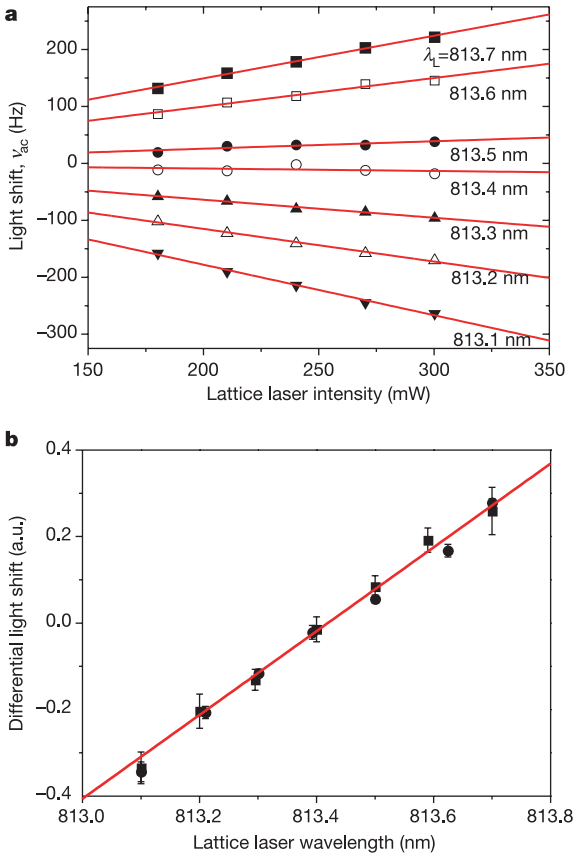


Figure 3 | Determination of the 'magic' wavelength. a, The light shift ν_{ac} of the clock transition was measured as a function of the lattice laser intensity I at several wavelengths λ_L near 813.5 nm. The differential light shift $\delta\nu_{ac}/\delta I$ was determined by the slope of the fitted lines. **b**, The differential light shift, which is proportional to the dipole polarizability difference $\Delta\alpha(\lambda_L)$, is plotted as a function of the lattice laser wavelength λ_L. The error bars represent the standard error. The 'magic' wavelength, where $\Delta\alpha(\lambda_L) = 0$, was determined to be $\lambda_L = 813.420(7)$ nm by a linear fit to the data points. a.u., arbitrary units.

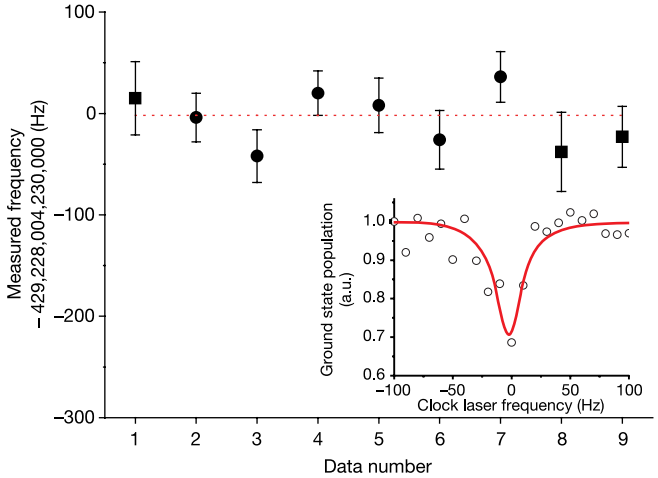


Figure 4 | Absolute frequency measurement of the $^1S_0 - {}^3P_0$ transition of Sr atoms in an optical lattice. The inset shows the typical clock transition resolving a linewidth of 27 Hz. The resonance frequency was measured by an optical frequency comb generator referenced to a commercial Cs clock. Each filled square/circle corresponds to a frequency measurement over 10^4 s. The standard deviation of the mean, which is typically 28 Hz, is shown by error bars. The filled circles correspond to measurements with an iodine-stabilized Nd:YAG laser as a flywheel, discussed in Methods. The 9 days of measurement determined the clock transition to be 429,228,004,229,952(15) Hz, applying the systematic correction of −45.7 Hz to the data shown in the figure.

interrogation time of 40 ms and a cycle time of 1 s, which is required for cooling, trapping and spectroscopy. We observed a nearly Fourier-limited linewidth of 27 Hz, which is the narrowest linewidth yet reported for neutral-atom-based optical clocks[7,8,11]. With 20 data points that took ~20 s to measure, the lorentzian fitting of the spectrum determined the centre frequency with an uncertainty of 2 Hz, corresponding to an Allan deviation $\sigma_y \approx 10^{-14}$ at $\tau = 1$ s. With the present measurement, the signal to noise ratio is limited by the shot-to-shot fluctuation (~10%) in the number of atoms loaded into the optical lattice. Normalization of the excitation rate[7], in addition to reduction of the clock laser linewidth[19], would improve the clock stability towards the projection noise limit for 10^6 atoms[1,7] of $\sigma_y(\tau) \approx 10^{-18}/\sqrt{\tau}$ (ref. 10), with τ measured in seconds.

The schematic for the absolute frequency measurement is shown in Fig. 2b. We repeated the above scan to determine the atomic resonance every 20 s and recorded the slow drift (~0.1 Hz s^{-1}) of the reference cavity that was used to prestabilize the clock laser. Simultaneously, we used the frequency comb, and measured the clock laser frequency against a commercial Cs clock (Agilent 5071A) that was calibrated with international atomic time (TAI) through GPS time[20]. These two sets of data were combined to deduce the absolute frequency of the lattice clock, as summarized in Fig. 4. In the figure, each filled square/circle represents a frequency measurement of typically 10^4 s to reduce the fractional instability of the Cs clock down to 6.5×10^{-14}, corresponding to an uncertainty of 28 Hz. With a total averaging time of $\tau \approx 9.4 \times 10^4$ s over 9 days, we reached an uncertainty of 9 Hz. The sources of the uncertainties and the corrections are listed in Table 1, and give a final value of 429,228,004,229,952(15) Hz, which is consistent with a measurement reported by Courtillot *et al.*[21] within their measurement uncertainty of 20 kHz. The collisional frequency shift, which was measured by varying the atom density by a factor of 3, was less than a measurement uncertainty (see Table 1). However, the shift may be present in this one-dimensional lattice configuration, as tens of atoms were typically trapped in a single lattice potential with an atom density of ~10^{12} cm^{-3}. This collision shift could ultimately be removed by spin-polarizing the ^{87}Sr atoms to activate the Fermi suppression of collisions[22], or by applying a three-dimensional lattice[9,10] with less than unity occupation, as shown in Fig. 1a.

This lattice clock scheme is applicable to other divalent atoms that have a hyperfine-mixing induced $J = 0 \rightarrow J = 0$ transition between long-lived states, such as ^{43}Ca (ref. 23), ^{171}Yb (ref. 24) and ^{199}Hg, in addition to the ^{87}Sr used here. A frequency comparison of lattice clocks operated with different atomic species would allow us to study the time variation of the fine structure constant[25]. The very high potential stability of this scheme would enable us to measure the fractional frequency difference at the 10^{-18} level, which corresponds to the gravitational red shift for a 1 cm height difference, in only one second. We expect that these features will have a great effect on future metrology, in areas ranging from fundamental physics to aspects of engineering, such as the remote sensing of natural resources.

METHODS

Optical frequency measurement. The development of ultrafast laser technology has revolutionized research on optical frequency measurement by introducing a precise optical frequency comb based on a femtosecond mode-locked laser[2]. The frequency of the nth comb component is expressed as $f_n = (n \times f_{rep}) + f_{CEO}$, where f_{rep} is the repetition rate of the laser pulse and f_{CEO} is the carrier-envelope offset frequency (Fig. 2b). The frequency comb obtained by injecting light from a mode-locked laser into a photonic crystal fibre[26] can cover more than one octave across the visible and near-infrared regions. f_{CEO} was measured directly with an

© 2005 Nature Publishing Group

f-to-$2f$ interferometer in the octave-spanning comb[3]. When f_{rep} and f_{CEO} are precisely controlled, the comb works as a 'frequency linker' that connects optical and radio frequencies. In the present experiment, a mode-locked Ti:sapphire laser with a repetition rate of 793 MHz was used to generate a frequency comb covering the 500–1,100 nm range. The frequency measurement was implemented by phase-locking the whole comb to the Sr clock laser and measuring f_{rep} against the Cs clock.

The frequency f_c of the clock laser prestabilized to a stable high-finesse reference cavity was frequency-scanned with an acousto-optic modulator (AOM) to find the atomic resonance frequency ν_0 every 20 s by observing a spectrum as shown in the inset of Fig. 4. As the frequency setting of the AOM $\delta(t) = f_c(t) - \nu_0$ measured the temporal drift of the clock laser, the atomic resonance frequency ν_0 could be determined by simultaneously measuring the clock-laser frequency $f_c(t)$. We observed a beat frequency $f_b = |f_c - f_n|$ between the clock laser and the nth comb component at 698 nm, and used it to servo-control the f_{rep} of the comb. Consequently, the clock-laser frequency $f_c(t)$ was converted to f_{rep}, which was then mixed down to about 7 MHz and recorded by a frequency counter with a gate time of 20 s. The clock-laser frequency $f_c(t)$ was calculated by using the formula $f_c = (n \times f_{rep}) + f_{CEO} \pm f_b$ with an integer $n \approx 5.4 \times 10^5$. The ambiguous sign was removed by observing the sign of the variation in the beat frequency f_b while f_{rep} was varied. We have thus obtained two series of data sets; the offset frequency $\{\delta(t)\}$ and the corresponding clock-laser frequency $\{f_c(t)\}$, both of which were determined every $t_m = 20$ s. The atomic resonance frequency was calculated as $\nu_0(k) = f_c(kt_m) - \delta(kt_m)$ for $k = 1, 2, \ldots, k_{max}$, where $k_{max} = T_{max}/t_m$ with T_{max} typically 10^4 s. Each filled square in Fig. 4 represents the mean value of $\{\nu_0(k)\}$, whose standard deviation of $\sigma \approx 620$ Hz was in agreement with the Cs clock's instability at $\tau = 20$ s. The standard deviation of the mean $\sigma/\sqrt{k_{max}}$ gave error bars of about 28 Hz.

In an alternative experimental scheme, a transportable iodine-stabilized Nd:YAG laser (NMIJ-Y1)[27] was used as a flywheel in the frequency measurement. The instability of the Nd:YAG laser $\sigma_y(\tau)$ is 2×10^{-13} for $\tau = 1$ s. This value improves after $\tau = 100$ s towards 3×10^{-14}, which is smaller than that of the Cs clock until $\tau = 10^5$ s. Furthermore, $\sigma_y(\tau)$ of the Nd:YAG laser is smaller than that of the clock laser for $\tau > 30$ s, which makes it possible to observe the slow drift of the clock laser. Because of the excellent short-term stability of the Nd:YAG laser, we also performed a frequency measurement by phase-locking the comb to the Nd:YAG laser. The clock-laser frequency $f_c(t)$ was measured by simultaneously introducing a second beat measurement between the clock laser and the comb component at 698 nm. The measurement results obtained with this scheme coincide with those obtained with the first measurement scheme, and are indicated by filled circles in Fig. 4.

The frequency of the Cs clock was monitored by the GPS time during the period of the frequency measurement (over one month). The timing difference between the 1 p.p.s. (pulse per second) signals generated by the Cs clock and a GPS disciplined oscillator was recorded continuously by a time interval counter. The frequency offset between the Cs clock and the GPS time was calculated from the recorded timing difference and found to be $-1.04(8) \times 10^{-13}$. The relationship between GPS time and TAI is available[20]. The agreement between the GPS time and the TAI was about one part in 10^{15} during the period of the frequency measurement.

Received 7 December 2004; accepted 10 March 2005.

1. Pereira Dos Santos, F. *et al.* Controlling the cold collision shift in high precision atomic interferometry. *Phys. Rev. Lett.* **89,** 233004 (2002).
2. Udem, Th., Reichert, J., Holzwarth, R. & Hänsch, T. W. Absolute optical frequency measurement of the cesium D_1 line with a mode-locked laser. *Phys. Rev. Lett.* **82,** 3568–3571 (1999).
3. Jones, D. J. *et al.* Carrier-envelope phase control of femtosecond mode-locked lasers and direct optical frequency synthesis. *Science* **288,** 635–639 (2000).
4. Diddams, S. A. *et al.* An optical clock based on a single trapped $^{199}Hg^+$ ion. *Science* **293,** 825–828 (2001).
5. Peik, E. *et al.* Limit on the present temporal variation of the fine structure constant. *Phys. Rev. Lett.* **93,** 170801 (2004).
6. Margolis, H. S. *et al.* Hertz-level measurement of the optical clock frequency in a single $^{88}Sr^+$ ion. *Science* **306,** 1355–1358 (2004).
7. Wilpers, G. *et al.* Optical clock with ultracold neutral atoms. *Phys. Rev. Lett.* **89,** 230801 (2002).
8. Oates, C. W., Curtis, E. A. & Hollberg, L. Improved short-term stability of optical frequency standards: approaching 1 Hz in 1 s with the Ca standard at 657 nm. *Opt. Lett.* **25,** 1603–1605 (2000).
9. Katori, H. in *Proc. 6th Symp. on Frequency Standards and Metrology* (ed. Gill, P.) 323–330 (World Scientific, Singapore, 2002).
10. Katori, H., Takamoto, M., Pal'chikov, V. G. & Ovsiannikov, V. D. Ultrastable optical clock with neutral atoms in an engineered light shift trap. *Phys. Rev. Lett.* **91,** 173005 (2003).
11. Ruschewitz, F. *et al.* Sub-kilohertz optical spectroscopy with a time domain atom interferometer. *Phys. Rev. Lett.* **80,** 3173–3176 (1998).
12. Allan, D. W. Statistics of atomic frequency standards. *Proc. IEEE* **54,** 221–230 (1966).
13. Dicke, R. H. The effect of collisions upon the Doppler width of spectral lines. *Phys. Rev.* **89,** 472–473 (1953).
14. Dehmelt, H. G. Mono-ion oscillator as potential ultimate laser frequency standard. *IEEE Trans. Instrum. Meas.* **31,** 83–87 (1982).
15. Jessen, P. S. & Deutsch, I. H. in *Advances in Atomic, Molecular and Optical Physics* **Vol. 37** (eds Bederson, P. & Walther, H.) 95–138 (Academic, San Diego, 1996).
16. Katori, H., Ido, T. & Kuwata-Gonokami, M. Optimal design of dipole potentials for efficient loading of Sr atoms. *J. Phys. Soc. Jpn* **68,** 2479–2482 (1999).
17. Ido, T. & Katori, H. Recoil-free spectroscopy of neutral Sr atoms in the Lamb-Dicke regime. *Phys. Rev. Lett.* **91,** 053001 (2003).
18. Takamoto, M. & Katori, H. Spectroscopy of the $^1S_0 - {}^3P_0$ clock transition of ^{87}Sr in an optical lattice. *Phys. Rev. Lett.* **91,** 223001 (2003).
19. Young, B. C., Cruz, F. C., Itano, W. M. & Bergquist, J. C. Visible lasers with subhertz linewidths. *Phys. Rev. Lett.* **82,** 3799–3802 (1999).
20. Bureau International des Poids et Mesures (BIPM), Circular T, No. 200 ⟨http://www1.bipm.org/en/scientific/tai/time_ftp.html⟩ (August 2004).
21. Courtillot, I. *et al.* Clock transition for a future optical frequency standard with trapped atoms. *Phys. Rev. A* **68,** 030501 (2003).
22. Gupta, S. *et al.* Radio-frequency spectroscopy of ultracold fermions. *Science* **300,** 1723–1726 (2003).
23. Degenhardt, C., Stoehr, H., Sterr, U., Riehle, F. & Lisdat, C. Wavelength-dependent ac Stark shift of the $^1S_0 - {}^3P_1$ transition at 657 nm in Ca. *Phys. Rev. A* **70,** 023414 (2004).
24. Porsev, S. G., Derevianko, A. & Fortson, E. N. Possibility of an optical clock using the $6^1S_0 \rightarrow 6^3P_0^0$ transition in $^{171,173}Yb$ atoms held in an optical lattice. *Phys. Rev. A* **69,** 021403 (2004).
25. Angstmann, E. J., Dzuba, V. A. & Flambaum, V. V. Relativistic effects in two valence-electron atoms and ions and the search for variation of the fine-structure constant. *Phys. Rev. A* **70,** 014102 (2004).
26. Knight, J. C. Photonic crystal fibres. *Nature* **424,** 847–851 (2003).
27. Hong, F.-L. *et al.* Frequency comparison of $^{127}I_2$-stabilized Nd:YAG lasers. *IEEE Trans. Instrum. Meas.* **48,** 532–536 (1999).
28. Ma, L.-S., Jungner, P., Ye, J. & Hall, J. L. Delivering the same optical frequency at two places: accurate cancellation of phase noise introduced by an optical fiber or other time-varying path. *Opt. Lett.* **19,** 1777–1779 (1994).
29. Rauschenbeutel, A., Schadwinkel, H., Gomer, V. & Meschede, D. Standing light fields for cold atoms with intrinsically stable and variable time phases. *Opt. Commun.* **148,** 45–48 (1998).

Acknowledgements We thank M. Yasuda, Y. Fukuyama and J. Jiang for assistance with the experiments, and A. Onae and S. Ohshima for discussions. This work was supported by the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communications of Japan.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.K. (katori@amo.t.u-tokyo.ac.jp).

© 2005 Nature Publishing Group

LETTERS

Micrometre-scale silicon electro-optic modulator

Qianfan Xu[1], Bradley Schmidt[1], Sameer Pradhan[1] & Michal Lipson[1]

Metal interconnections are expected to become the limiting factor for the performance of electronic systems as transistors continue to shrink in size. Replacing them by optical interconnections, at different levels ranging from rack-to-rack down to chip-to-chip and intra-chip interconnections, could provide the low power dissipation, low latencies and high bandwidths that are needed[1–4]. The implementation of optical interconnections relies on the development of micro-optical devices that are integrated with the microelectronics on chips. Recent demonstrations of silicon low-loss waveguides[5–7], light emitters[8], amplifiers[9–11] and lasers[12,13] approach this goal, but a small silicon electro-optic modulator with a size small enough for chip-scale integration has not yet been demonstrated. Here we experimentally demonstrate a high-speed electro-optical modulator in compact silicon structures. The modulator is based on a resonant light-confining structure that enhances the sensitivity of light to small changes in refractive index of the silicon and also enables high-speed operation. The modulator is 12 micrometres in diameter, three orders of magnitude smaller than previously demonstrated. Electro-optic modulators are one of the most critical components in optoelectronic integration, and decreasing their size may enable novel chip architectures.

Highly compact electro-optical modulators have been demonstrated in compound semiconductors[14]. However, in silicon, electro-optical modulation has been demonstrated only in large structures[15–17], and is therefore inappropriate for effective on-chip integration. Electro-optical control of light on silicon is challenging owing to its weak electro-optical properties[18]. The large dimensions of previously demonstrated structures were necessary to achieve a significant modulation of the transmission in spite of the small change of refractive index of silicon[15–17,19,20]. Liu *et al.* have recently demonstrated a high-speed silicon optical modulator based on a metal–oxide–semiconductor (MOS) configuration[17]. Their work showed a high-speed optical active device on silicon—a critical milestone towards optoelectronic integration on silicon. However, owing to the weak dependence of silicon's refractive index on the electron–hole pair concentration, and the limited mode confinement in the active region of the MOS device, the devices in ref. 17 have relatively large lengths (of the order of millimetres).

Light-confining resonating structures can enhance the effect of refractive index change on the transmission response[21–23]. In ref. 23, silicon ring resonators were used for all-optical modulation. The optical properties of the device were changed by using one beam of light to optically inject electrons and holes and thus control the flow of another beam of light. Using the principle of high confinement, here we present a silicon electro-optical modulator of a few micrometres in size[24].

The schematic of the electro-optic modulator is shown in Fig. 1. The modulator consists of a ring resonator coupled to a single waveguide. The transmission of the waveguide is highly sensitive to the signal wavelength and is greatly reduced at wavelengths in which the ring circumference corresponds to an integer number of guided wavelengths[23,25]. By tuning the effective index of the ring waveguide, the resonance wavelength is modified, which induces a strong modulation of the transmitted signal. The effective index of the ring is modulated electrically by injecting electrons and holes using a p-i-n junction embedded in the ring resonator. Figure 1 (inset) shows the cross-section of this rib waveguide. It consists of a strip waveguide formed on a 50-nm-thick slab layer. Because the thickness of the slab is much smaller than the wavelength propagating in the device (~1.5 μm), the mode profile of this waveguide is very close to that of the silicon strip waveguide. The highly doped p- and n-regions are defined around the ring using ion implantation. Ohmic contacts are deposited on the doped regions. To minimize absorption losses, the doped regions are formed approximately 1 μm away from the ring resonator, ensuring that the overlap of the resonating mode with the doped regions is minimal[26].

We fabricated the p-i-n ring resonator on a silicon-on-insulator substrate with a 3-μm-thick buried oxide layer. Both the waveguide coupling to the ring and that forming the ring have a width of 450 nm and a height of 250 nm. The diameter of the ring is 12 μm, and the spacing between the ring and the straight waveguide is 200 nm. To ensure high coupling efficiency between the waveguide and the incoming optical fibre, nanotapers[27] are fabricated at the ends of the waveguide. The structures are defined using electron-beam lithography followed by reactive ion plasma etching. Figure 2a shows the top-view scanning electron microscopy (SEM) image of the ring coupled to the waveguide with a close-up inset of the coupling region. Following the etching, the n+ and p+ regions of the diode are each defined with photolithography and implanted with phosphorus and boron to create concentrations of 10^{19} cm^{-3}. A

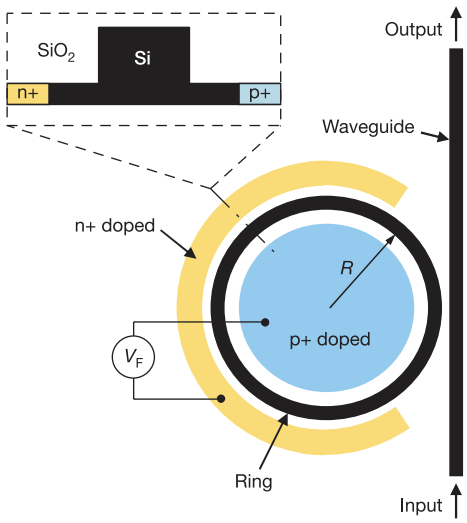


Figure 1 | Schematic layout of the ring resonator-based modulator. The inset shows the cross-section of the ring. R, radius of ring. V_F, voltage applied on the modulator.

[1]School of Electrical and Computer Engineering, Cornell University, 411 Phillips Hall, Ithaca, New York 14853, USA.

© 2005 Nature Publishing Group

1-μm-thick silicon dioxide layer is then deposited onto the wafer using plasma-enhanced chemical vapour deposition followed by an annealing process to activate the dopants (15 s at 1,050 °C for p+ and 15 min at 900 °C for n+). Holes were patterned using photolithography and then etched down to the doped silicon regions, followed by evaporation and liftoff of the titanium contacts. Figure 2b shows the top-view microscope image of the ring resonator after the metal contacts are formed.

Continuous-wave light (dominant electric field parallel to the plane of the substrate) from a tunable laser is coupled to the waveguide adjacent to the ring, and the output spectrum of the waveguide is measured. The free-spectrum range of the ring resonator is 15 nm. Figure 3 shows the relative transmission through the modulator around the resonance of 1,574 nm at different bias values of the p-i-n junction. The total optical loss of the testing sample, including the coupling loss between the fibre and the waveguide, and the propagation loss in the 7.5-mm-long waveguide leading to the modulator, is about 9 dB. The solid curve shows the spectrum when the forward bias (0.59 V) on the p-i-n junction is much lower than the built-in potential of the junction, and the current through the junction is below our detection limit (0.1 μA). The spectrum shows a 15-dB drop in transmission at the resonant wavelength of 1,574.9 nm. The low transmission at this wavelength is the initial condition of a 'dark-to-light' intensity modulator. The 3-dB bandwidth of the resonance $\Delta\lambda$ is 0.04 nm measured from the spectrum, corresponding to a quality factor, defined as $Q = \lambda/\Delta\lambda$, of 39,350. This Q factor corresponds to a cavity photon lifetime of $\tau_{cav} = \lambda^2/(2\pi c\Delta\lambda) = 33$ ps, where c is the speed of light in vacuum. Thus, despite the resonant nature of the structure, the photon confinement of the modulator is not limiting its speed. The small ripples (~1 dB) on the waveform originate from the reflections at both ends of the waveguide, which can be eliminated by anti-reflection coating.

The electron–hole pair density in the cavity increases as the forward bias on the p-i-n junction increases. The dashed and dotted curves in Fig. 3 show the spectra when the bias voltage is 0.87 V and 0.94 V, and the current is 11.1 μA and 19.9 μA, respectively. In both cases the resonance is blue-shifted owing to the lowering of the effective index caused by the increase of the electron–hole pair density in the cavity[18]. The depths of the notches in the spectra decrease owing to the increased optical absorption in the ring induced by the electrons and holes. The extra absorption per round-trip in the ring is estimated from the spectra to be 0.03 dB at 0.87 V bias and 0.05 dB at 0.94 V bias. At the probe wavelength of 1,573.9 nm, we can see from Fig. 3 that 97% (15 dB) modulation can be obtained with less than 0.3-V bias voltage change. At this wavelength, because light does not couple to the ring when the electrons and holes are generated, the effect of the extra absorption in the ring on the modulation is less than 0.1 dB.

The light-confining nature of the modulator not only enables shrinking of the device size, but also enables high-speed operation under p-i-n configuration. The p-i-n configuration of the modulator, as opposed to the MOS configuration[17], is important for achieving high modulation depth, because the overlap between the region where the index is changed and the optical mode of the waveguide is large. However, p-i-n devices have traditionally been considered as relatively slow compared with MOS devices. In p-i-n devices, while extraction of electrons and holes in reverse-biased operation can be fast, down to tens of picoseconds[28], injection of electrons and holes in forward-bias operation is slow, which limits the rise time of the p-i-n to of the order of 10 ns, as measured in our device. The resonating nature of the modulator removes this speed limitation.

The inset in Fig. 3 shows the transfer function of the device, that is, the transmission of the modulator at the wavelength of 1,573.9 nm with different bias voltages. This transfer function shows that the resonating nature of the device enables voltages larger than 0.9 V to be applied without modifying the optical transmission ($T \approx 1$). This is because at these voltages, the resonance of the device is completely detuned from the probe wavelength. This insensitivity of the optical transmission at higher bias voltage is in strong contrast to Mach–Zehnder modulators, in which higher voltage strongly affects the transmission. When the device is operated at higher voltages, the optical transmission can reach ~1 well before the p-i-n junction reaches its steady state. This means that the optical rise time can be far less than the electrical rise time of ~10 ns with high forward bias, which is crucial for achieving high-speed modulation.

To measure the dynamic response of the modulator, electrical signals generated by a pulse-pattern generator and a wideband microwave amplifier are used to drive a modulator. The output of the waveguide is sent to a 12-GHz detector and the waveform is recorded on an oscilloscope. Figure 4a, b shows the waveform of both the driving signal using a non-return-to-zero (non-RZ) pattern and the optical output, demonstrating high modulation depths at 0.4 Gbit s^{-1}. The peak-to-peak voltage (V_{pp}) of the driving signal is 3.3 V, from −1.85 V to +1.45 V. The speed of the device is limited by the fact that the p-i-n junction is formed on only part of the ring resonator (see Fig. 1). During the forward-biasing period, electrons and holes diffuse into the section of the ring that is not part of the p-i-n junction (the section close to the straight waveguide),

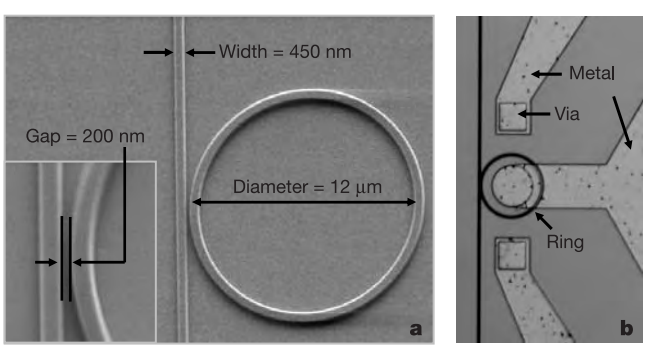


Figure 2 | SEM and microscope images of the fabricated device. a, Top-view SEM image of the ring coupled to the waveguide with a close-up view of the coupling region. **b**, Top-view microscope image of the ring resonator after the metal contacts are formed. The metal contact on the central p-doped region of the ring goes over the ring with a 1-μm-thick silicon dioxide layer between the metal and the ring.

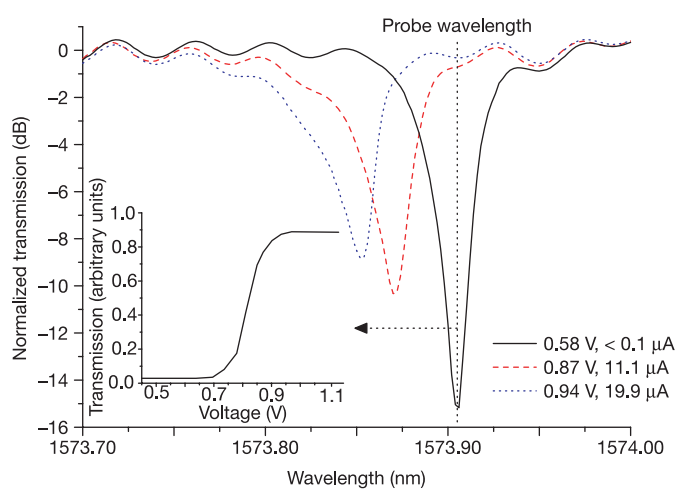


Figure 3 | DC measurement of the ring resonator. The main panel shows the transmission spectra of the ring resonator at the bias voltages of 0.58 V, 0.87 V, and 0.94 V, respectively. The vertical dashed line marks the position of the probe wavelength used in the transfer function and dynamic modulation measurements. The inset shows the transfer function of the modulator for light with a wavelength of 1,573.9 nm.

© 2005 Nature Publishing Group

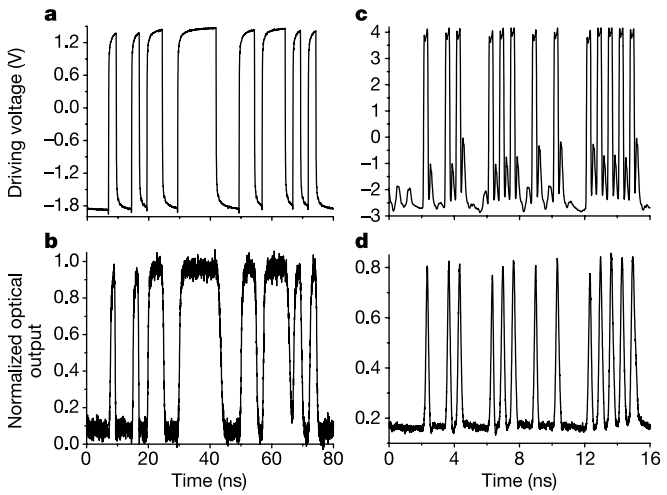


Figure 4 | Waveforms of the electrical driving signal and the transmitted optical signal. a, Waveform of the 32-bit pseudo-random non-RZ signal at 0.4 Gbit s^{-1} applied on the modulator. **b**, Waveform of the output optical power of the modulator with non-RZ modulation, normalized to the output power when the ring resonator is completely out of resonance. **c**, Waveform of the 127-bit pseudo-random bit sequence RZ signal at 1.5 Gbit s^{-1} applied on the modulator with a microwave amplifier. **d**, Waveform of the normalized output optical power of the modulator with RZ modulation.

where they cannot be efficiently extracted during the reverse-biased period, leading to a longer fall time (~1.5 ns) following long forward-biasing periods. This longer fall time is determined by the intrinsic lifetime of the electrons and holes in the non-p-i-n region and is independent of the reverse-biased voltage.

Figure 4c, d shows high modulation depths of the device at 1.5 Gbit s^{-1}. The modulator is driven by a 1.5 Gbit s^{-1} RZ signal with $V_{pp} = 6.9$ V (from −2.8 V to +4.1 V). The RZ signal does not have a long forward-biasing period, so the diffusion of electrons and holes into the non-p-i-n region is reduced. This enables the device to work at a higher bit-rate, despite the fact that the RZ signal generally requires twice as much bandwidth as the non-RZ signal for the same bit-rate. The higher voltage swing allows faster injection and extraction of electrons and holes, and the rise time and fall time of the optical output are found to be 200 ps and 150 ps, respectively. Note that, owing to the step-like transfer function of the modulator, the optical output does not follow the distortion of the electrical signal caused by reflections in the electrical driving system.

The driving voltage of the modulator can be decreased by reducing the contact resistance (currently ~4 kΩ). Using a typical value of ~100 Ω (ref. 29), the voltage drop on the contact resistance, estimated to be ~4 V_{pp} in the RZ-driven experiment, can be reduced to ~0.1 V_{pp}. We estimate that the silicon ring resonator, with the p-i-n junction covering the whole ring and low contact resistance, can be modulated at a bit-rate higher than 5 Gbit s^{-1} with a non-RZ signal of ~3 V_{pp}.

The wavelength-selective modulation property of the modulator can be used for building wavelength division multiplexing (WDM) interconnections, which can greatly extend the bandwidth of optical interconnections. Given the short length of the modulator (<20 μm) and the waveguide propagation loss of 4 ± 1 dB cm^{-1} (refs 6, 7), the insertion loss of the modulator is negligible to light with wavelength detuned from the ring resonance. The small insertion loss of the modulator makes it possible to cascade multiple modulators along a single waveguide and modulate each WDM channel independently. Note that this implementation would require the dimensions of the ring resonators to be controlled such that the resonant wavelength of each ring resonator equals the wavelength of the corresponding WDM channel.

Received 8 February; accepted 22 March 2005.

1. Miller, D. A. B. Optical interconnects to silicon. *IEEE J. Sel. Top. Quant. Electron.* **6,** 1312–1317 (2000).
2. Meindl, J. D. *et al.* Interconnect opportunities for gigascale integration. *IBM Res. Dev.* **46,** 245–263 (2002).
3. Mohammed, E. M. *et al.* Optical I/O technology for digital VLSI. *Proc. SPIE* **5358,** 60–70 (2004).
4. Kibar, O., Van Blerkom, D. A., Fan, C. & Esener, S. C. Power minimization and technology comparisons for digital free-space optoelectronic interconnections. *J. Lightwave Technol.* **17,** 546–555 (1999).
5. Lee, K. K., Lim, D. R. & Kimerling, L. C. Fabrication of ultralow-loss Si/SiO_2 waveguides by roughness reduction. *Opt. Lett.* **26,** 1888–1890 (2001).
6. Vlasov, Y. A. & McNab, S. J. Losses in single-mode silicon-on-insulator strip waveguides and bends. *Opt. Express* **12,** 1622–1631 (2004).
7. Dumon, P. *et al.* Low-loss SOI photonic wires and ring resonators fabricated with deep UV lithography. *IEEE Photon. Technol. Lett.* **16,** 1328–1330 (2004).
8. Chan, S. & Fauchet, P. M. Silicon microcavity light emitting devices. *Opt. Mater.* **17,** 31–34 (2001).
9. Jones, R. *et al.* Net continuous wave optical gain in a low loss silicon-on-insulator waveguide by stimulated Raman scattering. *Opt. Express* **13,** 519–525 (2005).
10. Espinola, R. L., Dadap, J. I., Osgood, R. M., McNab, S. J. & Vlasov, Y. A. Raman amplification in ultrasmall silicon-on-insulator wire waveguides. *Opt. Express* **12,** 3713–3718 (2004).
11. Xu, Q., Almeida, V. R. & Lipson, M. Time-resolved study of Raman gain in highly confined silicon-on-insulator waveguides. *Opt. Express* **12,** 4437–4442 (2004).
12. Boyraz, O. & Jalai, B. Demonstration of a silicon Raman laser. *Opt. Express* **12,** 5269–5273 (2004).
13. Rong, H. *et al.* An all-silicon Raman laser. *Nature* **433,** 292–294 (2005).
14. Sadagopan, T., Choi, S. J., Dapkus, P. D. & Bond, A. E. *Digest of the LEOS Summer Topical Meetings* MC2–3 (IEEE, Piscataway, New Jersey, 2004).
15. Tang, C. K. & Reed, G. T. Highly efficient optical phase modulator in SOI waveguides. *Electron. Lett.* **31,** 451–452 (1995).
16. Dainesi, P. *et al.* CMOS compatible fully integrated Mach–Zehnder interferometer in SOI technology. *IEEE Photon. Technol. Lett.* **12,** 660–662 (2000).
17. Liu, A. *et al.* A high-speed silicon optical modulator based on a metal-oxide-semiconductor capacitor. *Nature* **427,** 615–618 (2004).
18. Soref, R. A. & Bennett, B. R. Electrooptical effects in silicon. *IEEE J. Quant. Electron.* **23,** 123–129 (1987).
19. Gan, F. & Kärtner, F. X. High-speed silicon electro-optic modulator design. *IEEE Photon. Technol. Lett.* (in the press).
20. Sciuto, A., Libertino, S., Alessandria, A., Coffa, S. & Coppola, G. Design, fabrication, and testing of an integrated Si-based light modulator. *J. Lightwave Technol.* **21,** 228–235 (2003).
21. Heebner, J. E. *et al.* Enhanced linear and nonlinear optical phase response of AlGaAs microring resonators. *Opt. Lett.* **29,** 769–771 (2004).
22. Maune, B., Lawson, R., Gunn, C., Scherer, A. & Dalton, L. Electrically tunable ring resonators incorporating nematic liquid crystals as cladding layers. *Appl. Phys. Lett.* **83,** 4689–4691 (2003).
23. Almeida, V. R., Barrios, C. A., Panepucci, R. R. & Lipson, M. All-optical control of light on a silicon chip. *Nature* **431,** 1081–1084 (2004).
24. Pradhan, S., Almeida, V. R., Barrios, C. & Lipson, M. *Optical Amplifiers and Their Applications / Integrated Photonics Research Topical Meetings* IWA5 CD-ROM (The Optical Society of America, Washington, DC, 2004).
25. Little, B. E. *et al.* Ultra-compact Si-SiO_2 microring resonator optical channel dropping filters. *IEEE Photon. Technol. Lett.* **10,** 549–551 (1998).
26. Barrios, C. A., Almeida, V. R., Panepucci, R. & Lipson, M. Electrooptic modulation of silicon-on-insulator submicrometer-size waveguide devices modulator. *J. Lightwave Technol.* **21,** 2232–2239 (2003).
27. Almeida, V. R., Panepucci, R. R. & Lipson, M. Nanotaper for compact mode conversion. *Opt. Lett.* **28,** 1302–1304 (2003).
28. Muller, J. Thin silicon film p-i-n photodiodes with internal reflection. *IEEE J. Solid-State Circ.* **13,** 173–179 (1978).
29. Faith, T. J., Irven, R. S., Plante, S. K. & O'Neill, J. J. Contact resistance: Al and Al-Si to diffused N+ and P+ silicon. *J. Vac. Sci. Technol. A* **1,** 443–448 (1983).

Acknowledgements This work has been partially carried out as part of the Interconnect Focus Center Research Program at Cornell University, supported in part by the Microelectronics Advanced Research Corporation (MARCO), its participating companies, and DARPA. We acknowledge support by the National Science Foundation (NSF). We thank G. Pomrenke, AFOSR, for supporting the work. We also acknowledge support by the Cornell Center for Nanoscale Systems. The devices were fabricated at the Cornell Nano-Scale Science & Technology Facility. We also thank J. Shah, DARPA, for funding this work as part of the EPIC programme.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.L. (lipson@ece.cornell.edu).

© 2005 Nature Publishing Group

LETTERS

Real-time forecasts of tomorrow's earthquakes in California

Matthew C. Gerstenberger[1], Stefan Wiemer[2], Lucile M. Jones[1] & Paul A. Reasenberg[3]

Despite a lack of reliable deterministic earthquake precursors, seismologists have significant predictive information about earthquake activity from an increasingly accurate understanding of the clustering properties of earthquakes[1–4]. In the past 15 years, time-dependent earthquake probabilities based on a generic short-term clustering model have been made publicly available in near-real time during major earthquake sequences. These forecasts describe the probability and number of events that are, on average, likely to occur following a mainshock of a given magnitude, but are not tailored to the particular sequence at hand and contain no information about the likely locations of the aftershocks. Our model builds upon the basic principles of this generic forecast model in two ways: it recasts the forecast in terms of the probability of strong ground shaking, and it combines an existing time-independent earthquake occurrence model based on fault data and historical earthquakes[5] with increasingly complex models describing the local time-dependent earthquake clustering[1,2]. The result is a time-dependent map showing the probability of strong shaking anywhere in California within the next 24 hours. The seismic hazard modelling approach we describe provides a better understanding of time-dependent earthquake hazard, and increases its usefulness for the public, emergency planners and the media.

We calculate the time-dependent seismic hazard by combining stochastic models derived from the Gutenberg–Richter relationship[6] and the modified Omori law[7]: starting from a background model of the long-term, spatially varying Poisson hazard[5], we add clustering models of varying complexity depending on the amount of currently available aftershock data. Most of the time and over most of California, a recent earthquake large enough to generate any appreciable increase in hazard will not have occurred and we then assume that the background model adequately represents the hazard. After an event of local magnitude $M_1 = 3$ or larger, the additional hazard associated with possible aftershocks[2] is added to the background hazard until this time-dependent contribution decays below the background level. We do not limit an aftershock to a magnitude smaller than its preceding mainshock[8], so that a foreshock is simply an event whose aftershock happens to be bigger than itself; this permits the use of the same statistical description for foreshock and aftershock hazard.

Our clustering model is a composite of up to three models of increasing complexity: a generic-clustering model derived from the original Reasenberg and Jones model[1,2], a sequence-specific model, and a spatially heterogeneous model. In the generic clustering model, the rate at time t of aftershocks with magnitude $M \geq M_c$ is given by:

$$\lambda(t) = 10^{a' + b(M_m - M)} / (t + c)^p$$

where a', b, c and p are constants and M_m is the mainshock magnitude. If the ongoing earthquake sequence produces a sufficient number of aftershocks, a sequence-specific model is calculated in addition to the generic model; this model uses the a posteriori parameter values ($\hat{a}'$, $\hat{b}$ and $\hat{p}$) estimated for the sequence. Concurrent with the sequence-specific model, a third, spatially heterogeneous[9,10] model is calculated, whose parameters (a'_i, b_i and p_i) are calculated independently at the nodes of a 5-km grid covering the region, with parameter estimates for each node based on the seismicity within 15 km of it. However, for any given node, it is not always possible to make parameter estimates owing to the need for a minimum number of events to obtain a robust estimate of the parameters. See the online Supplementary Methods section for a full treatment of the calculations.

We do not limit an aftershock to occur at the same location as its mainshock. In the first two models, the mainshock is assumed initially to be a point source. This is modified to be a finite source, if and when the geometry of the source is known. The total rate of aftershocks is determined by the model and is distributed across an area that extends one-half of a fault length from the source. The spatial density of aftershocks is assumed to be proportional to $1/r^2$, where r is the horizontal distance from the mainshock's source. In the spatially heterogeneous model, the spatial variations in the rate are calculated directly from the distribution of aftershocks that have already occurred.

We calculate the expected rate of $4 \leq M \leq 8$ aftershocks at each node and for each model and combine the rates using Akaike weights, based on the corrected Akaike Information Criterion (AICc)[11]. The weights are based on the model's likelihood score, the amount of data and the number of free parameters that must be estimated, thus ensuring a gradual transition from the generic-clustering to the sequence-specific and spatially heterogeneous model as sufficient data warranting a more complex model become available. Because the AICc prefers fewer free parameters, the spatially heterogeneous model will be strongly weighted only in active areas.

The hazard of earthquake shaking at any site depends on the expected rate of earthquake sources nearby, their magnitudes and distances from the site, propagation effects (primarily geometric spreading and seismic attenuation) and local soil conditions at the site. We use an empirical attenuation function[12] to estimate peak ground acceleration (PGA) from a possible aftershock. We currently ignore local site effects and convert the PGA to an equivalent Modified Mercalli Intensity (MMI)[13]. The seismic hazard at a location is expressed as the probability that ground shaking exceeding MMI VI will occur there in the next 24 hours. We assume an isotropic distribution of shaking from any future earthquake as aftershocks and foreshocks do not have to have the same focal mechanism as their mainshock.

A snapshot of the probability of exceeding MMI VI in California on 28 July 2004 is shown in Fig. 1c. This is a composite of the time-independent background from geological information and

[1]US Geological Survey, 525 S. Wilson Ave, Pasadena, California 91106, USA. [2]ETH Zürich, Institute of Geophysics, 8093 Zürich, Switzerland. [3]US Geological Survey, 345 Middlefield Road, Menlo Park, California 94025, USA.

© 2005 Nature Publishing Group

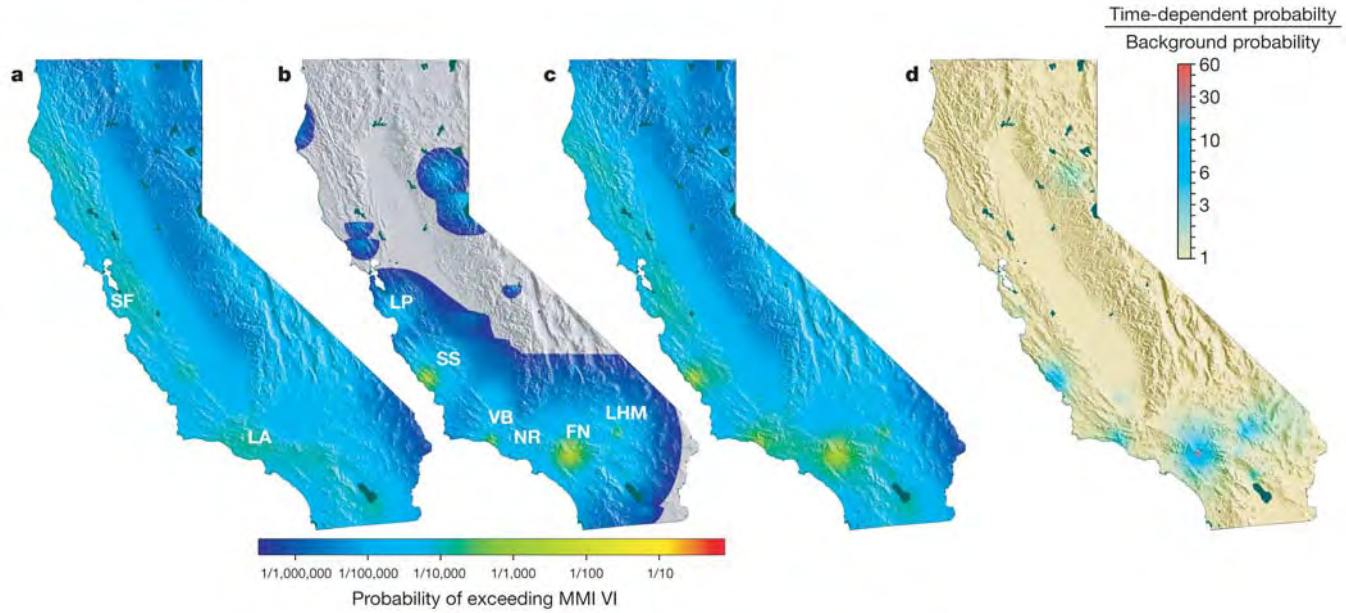


Figure 1 | Maps of California, showing the probability of exceeding MMI VI over the next 24-h period. The period starts at 14:07 Pacific Daylight Time on 28 July 2004. **a**, The time-independent hazard based on the 1996 USGS hazard maps for California. SF and LA are the locations of San Francisco and Los Angeles, respectively. **b**, The time-dependent hazard which exceeds the background including contributions from several events: 22 December 2003, San Simeon (SS, $M_w = 6.5$), a $M_l = 4.3$ earthquake 4 days earlier near Ventura (VB), a $M_l = 3.8$ event 30 min before the map was made near San Bernardino (FN), the 1999 Hector Mine $M_w = 7.1$ earthquake in the Mojave desert (LHM), and the 1989 $M_w = 6.9$ Loma Prieta (LP) earthquake. **c**, The combination of these two contributions, representing the total forecast of the likelihood of ground shaking in the next 24-hour period. **d**, The ratio of the time-dependent contribution to the background.

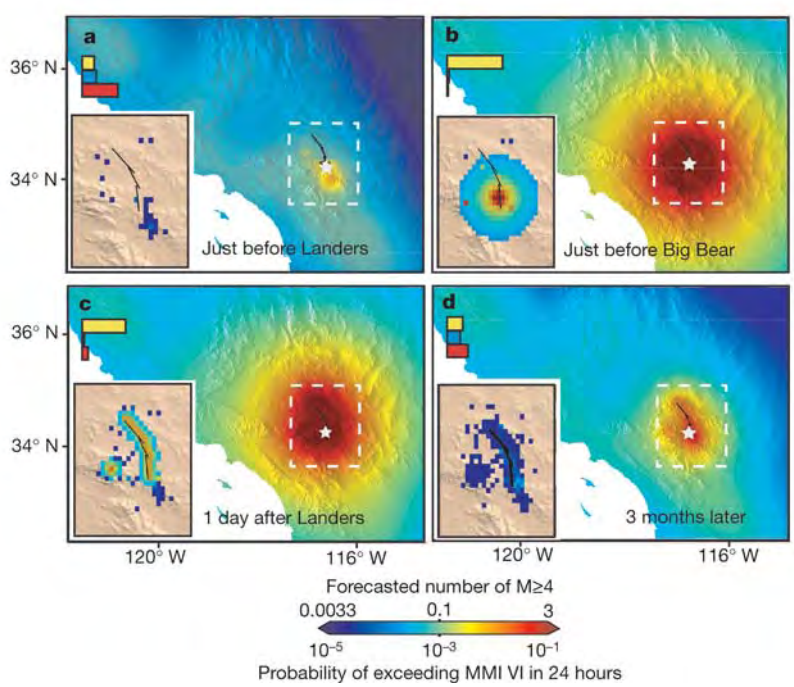


Figure 2 | Hazard maps calculated for four 24-hour intervals in the vicinity of the 1992 Landers ($M = 7.3$) earthquake. The colour shows the probability of exceeding MMI VI shaking intensity during interval (labels below colour scale). The black line indicates the Landers rupture and the white star is the Landers epicentre. The inset maps correspond to area within dashed white lines and indicate expected number of $M \geq 4.0$ aftershocks in the 24-hour interval (labels above colour scale). Histograms on upper left of figures indicate relative mean weight over inset area given to generic model (yellow), sequence-specific model (blue) and spatially heterogeneous model (red). **a**, Interval beginning 1 min before Landers mainshock shows increased hazard south of the Landers epicentre associated with the foreshocks to Landers (largest $M_l = 3.0$) and aftershocks of the ($M_w = 6.1$) Joshua Tree earthquake two months earlier. **b**, Interval beginning 2 hours after the mainshock and 1 hour before the ($M_w = 6.3$) Big Bear aftershock. **c**, Interval beginning 24 h after Landers mainshock. **d**, Interval beginning three months after mainshock.

© 2005 Nature Publishing Group

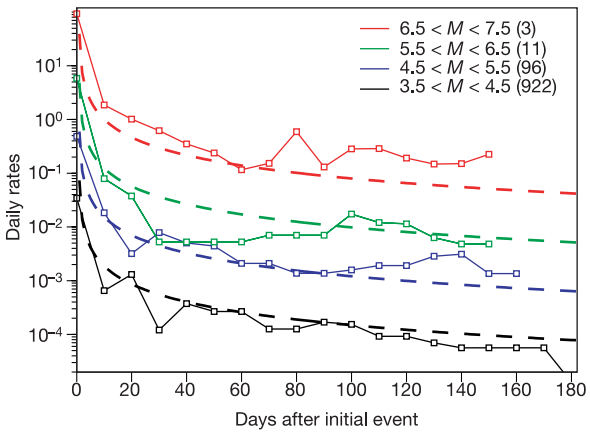


Figure 3 | Calculated and observed rates of events $M \geq 4$ in 24-hour intervals following mainshocks occurring between 1988 and 2002 in southern California. Dashed lines show the rates forecasted by the generic California clustering model (without cascades) for the mainshock magnitude (M) shown. For this test a simple circular aftershock zone implementation (solid lines) gives the observed rates of $M \geq 4.0$ aftershocks following all mainshocks with magnitude within 0.5 units of M. The aftershock zones are defined as the areas within one rupture length of the mainshock epicentre.

historical earthquake data (Fig. 1a) and the time-dependent contributions (Fig. 1b). The ratio of the time-dependent contribution to the background probability is shown in Fig. 1d.

We treat every earthquake ($M \geq 3$) as a potential mainshock and calculate an aftershock model for it. At any time, more than one earthquake may be contributing to the expected rate of aftershocks at a given location, most commonly after a large aftershock. In these cases, the earthquake producing the highest expected rate at a location is adopted, and its rate assigned to the node. Thus, a cascade of aftershock sequences[3] can be modelled, with secondary and tertiary sequences represented where and when they exceed the rate of the primary sequence. Having translated all models into common units of daily rate of events for each magnitude bin at each location, we have a straightforward procedure for deciding how to terminate a cluster: in our definition, a sequence remains active as long as at any one location a higher number of events ($M \geq 4$) is forecast than is contained in the background model at the same location.

The application of our approach to the 1992 Landers sequence is demonstrated in Fig. 2. Just before the mainshock, the background model dominates the earthquake hazard, except in the Landers region where foreshocks ($M_1 \leq 3.0$) had just occurred and near the moment magnitude (M_w) = 6.1 Joshua Tree earthquake that occurred two months earlier (Fig. 2a). In the first hours after the mainshock, the generic clustering model dominates and is isotropic relative to the mainshock's epicentre (Fig. 2b). By one day after the mainshock, a mainshock rupture model is available, allowing fault-specific versions of both the generic and sequence-specific models to be calculated (Fig. 2c). Also, the $M_w = 6.3$ Big Bear aftershock has occurred, and is reflected in the forecast rate of aftershocks. However, the hazard is still nearly isotropic owing to the smoothing inherent in the hazard calculation and the shape of the combined aftershock zones. Three months later (Fig. 2d), the spatially heterogeneous model is used in most nodes. The modelled rate of aftershocks is lower but still heterogeneous, producing a heterogeneous hazard forecast.

As a first test, we verified that the generic clustering model[1,2] describes the average clustering activity of California reasonably well. Using data from 1988–2002, after the period used to initially develop the model and thus independent data, we compute the average daily rate of events following an earthquake of a given size (Fig. 3).

Next, using a likelihood-based testing algorithm developed as part of the Regional Earthquake Likelihood Models group of the Southern California Earthquake Centre (D. Schorlemmer, M.C.G., D. J. Jackson & S.W.; manuscript in preparation) we compared our model against successively more challenging null hypotheses. First, we showed that the background model of the 1996 National Hazard map could be rejected with a significance of <1%. This is hardly surprising, as this model does not include earthquake clustering. In the second and third tests, we compared the full composite model to stages of the model, the second using only the generic clustering model, and the third using sequence-specific parameters when they could be determined. Again, we were able to reject both partial models with <1% significance. We can thus state that the composite model did a better job of predicting the seismicity of 1992–2001 in the region of Fig. 2 than the National Hazard maps, an average clustering model, or a sequence-specific parameter model.

Many earthquake-forecasting case studies have been made[14], but few have been tested systematically against a null hypothesis[15]. We believe strongly that quantitative testing is key to making progress in earthquake forecasting research. Our time-dependent hazard model has been tested against the null hypothesis of the long-term hazard model and represents our empirical understanding of aftershock behaviour. It could be a null hypothesis for future statistical or physical models in time-dependent forecast-related research.

The first hours to days after a mainshock are the most hazardous time, with the daily probabilities of exceeding MMI VI that approach 1 after the largest possible mainshocks. The need to characterize the relative and absolute increase in hazard is therefore at its greatest when emergency response activities are underway. Because the hazard is highest immediately after the mainshock, it is important to update quickly, preferably in an automatic system. Reliable real-time monitoring of larger aftershocks sequences is now possible in California, as demonstrated during the 1999 $M_w = 7.1$ Hector Mine event[16]; however, further improvements in monitoring and processing are likely once the new Advanced National Seismic System (ANSS) is implemented. This should enhance our model's forecasting ability because we will be able to move more quickly from the isotropic to a more realistic, heterogeneous model. Similar to real-time seismicity maps and ShakeMaps[13], 24-hour forecast maps such as the one shown in Fig. 1 are now available through the web (http://pasadena.wr.usgs.gov/step).

Received 19 January; accepted 7 April 2005.

1. Reasenberg, P. A. & Jones, L. M. Earthquake aftershocks: Update. *Science* **265**, 1251–1252 (1994).
2. Reasenberg, P. A. & Jones, L. M. Earthquake hazard after a mainshock in California. *Science* **243**, 1173–1176 (1989).
3. Helmstetter, A. & Sornette, D. Sub-critical and supercritical regimes in epidemic models of aftershock sequences, D. *J. Geophys. Res.* **107**, 2237, doi:10.1029/2001JB001580 (2002).
4. Felzer, K. R., Abercrombie, R. E. & Ekstrom, G. A common origin for aftershocks and multiplets. *Bull. Seismol. Soc. Am.* **94**, 88–99 (2004).
5. Frankel, A. *et al.* National seismic hazard maps. *U. S. Geological Survey Open-File Report*, 96–532 (1996).
6. Gutenberg, B. & Richter, C. F. Magnitude and energy of earthquakes. *Ann. Geofis.* **9**, 1–15 (1954).
7. Utsu, T., Ogata, Y. & Matsu'ura, R. S. The centenary of the Omori formula for a decay law of aftershock activity. *J. Phys. Earth* **43**, 1–33 (1995).
8. Reasenberg, P. A. Foreshock occurrence before large earthquakes. *J. Geophys. Res.* **104**, 4755–4768 (1999).
9. Wiemer, S., Gerstenberger, M. & Hauksson, E. Properties of the aftershock sequence of the 1999 Mw 7.1 Hector Mine earthquake: implications for aftershock hazard. *Bull. Seismol. Soc. Am.* **92**, 1227–1240 (2002).
10. Wiemer, S. & Katsumata, K. Spatial variability of seismicity parameters in aftershock zones. *J. Geophys. Res.* **104**, 13135–13151 (1999).
11. Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference A Practical Information-Theoretic Approach* 488 (Springer, New York, 2002).
12. Boore, D. M., Joyner, W. B. & Fumal, T. E. Equations for estimating horizontal response spectra and peak acceleration from western North America earthquakes: A summary of recent work. *Seismol. Res. Lett.* **68**, 128–153 (1997).
13. Wald, D. J., Quitoriano, V. & Heaton, T. H. TriNet "ShakeMaps": rapid generation of instrumental ground motion and intensity maps for earthquakes

© 2005 Nature Publishing Group

in southern California. *Earthquake Spectra* **15,** 537–556 (1999).
14. Toda, S., Stein, R. S., Reasenberg, P. A. & Dieterich, J. H. Stress transferred by the Mw = 6.9 Kobe, Japan, shock: effect on aftershocks and future earthquake probabilities. *J. Geophys. Res.* **103,** 24543–24565 (1998).
15. Geller, R. J. Earthquake prediction: a critical review. *Geophys. J. Int.* **131,** 425–450 (1997).
16. Wiemer, S. Introducing probabilistic aftershock hazard mapping. *Geophys. Res. Lett.* **27,** 3405–3408 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are very grateful for the reviews from E. Field, D. Jackson, R. Simpson and A. Cornell. The work was supported by the Southern California Earthquake Center, ETH Zurich and the US Geological Survey.

Author Contributions This work was completed as part of the PhD thesis of M.C.G. under the supervision of S.W. at ETH-Zürich. L.M.J. and P.A.R. provided significant contributions to the model development.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.C.G. (mattg@usgs.gov).

© 2005 Nature Publishing Group

LETTERS

Direct dating of Early Upper Palaeolithic human remains from Mladeč

Eva M. Wild[1], Maria Teschler-Nicola[3], Walter Kutschera[1], Peter Steier[1], Erik Trinkaus[4] & Wolfgang Wanek[2]

The human fossil assemblage from the Mladeč Caves in Moravia (Czech Republic)[1] has been considered to derive from a middle or later phase of the Central European Aurignacian period on the basis of archaeological remains (a few stone artefacts and organic items such as bone points, awls, perforated teeth)[2], despite questions[3] of association between the human fossils and the archaeological materials and concerning the chronological implications of the limited archaeological remains[4]. The morphological variability in the human assemblage, the presence of apparently archaic features in some specimens, and the assumed early date of the remains have made this fossil assemblage pivotal in assessments of modern human emergence within Europe[5–7]. We present here the first successful direct accelerator mass spectrometry radiocarbon dating of five representative human fossils from the site. We selected sample materials from teeth and from one bone for ^{14}C dating. The four tooth samples yielded uncalibrated ages of ~31,000 ^{14}C years before present, and the bone sample (an ulna) provided an uncertain more-recent age. These data are sufficient to confirm that the Mladeč human assemblage is the oldest cranial, dental and postcranial assemblage of early modern humans in Europe and is therefore central to discussions of modern human emergence in the northwestern Old World and the fate of the Neanderthals.

The Mladeč site has significance for both human evolutionary and archaeological issues[3,8,9], and the relevance of its remains has increased as a result of the recent dating of the purportedly Aurignacian-age modern human remains from Velika Pećina (Croatia)[10], Hahnöfersand (Germany)[11] and Vogelherd (Germany)[3] to the Holocene epoch, the remains from Koněprusy (Czech Republic)[9] to the Magdalenian period, and those from Cro-Magnon (France)[12] and La Rochette (France)[13] to the Gravettian period. The only directly dated European modern human fossils of Aurignacian age are the Peştera cu Oase (Romania) mandible and cranium at ~35,000 ^{14}C years before present (that is, ~35 ^{14}C kyr BP)[14], the Kent's Cavern (UK) maxilla at ~31 ^{14}C kyr BP[15], the Peştera Muierii (Romania) remains at ~30 ^{14}C kyr BP[16], and the Peştera Cioclovina (Romania) cranium at ~29 ^{14}C kyr BP[16], none of which has a secure and diagnostic archaeological association. Moreover, at least the Oase fossils overlap in time with late Neanderthals from for example, Vindija (Croatia), which is at present dated to ~29 ^{14}C kyr BP[10] and Arcy-sur-Cure (France) at ~34 ^{14}C kyr BP[17]. The assessment of whether the Mladeč fossils are indeed Aurignacian in age, and if so, their chronological position within the Aurignacian time span, has become central to understanding early modern humans in Europe.

The Mladeč human remains are universally accepted as those of early modern humans since the analysis of Szombathy[1]. However, there is an ongoing debate as to whether they exhibit distinctive archaic features, indicative of some degree of regional Neanderthal ancestry, or are morphologically aligned solely with recent humans and therefore document only a dispersal of modern humans into Europe. The purportedly archaic, or Neanderthal, features include aspects of the sagittal cranial profile and robust supraorbital regions

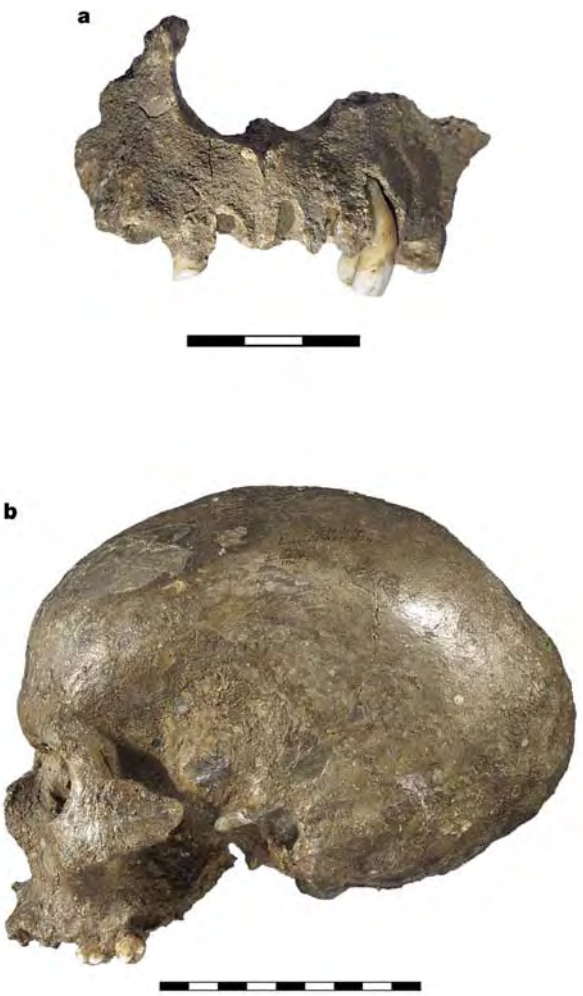


Figure 1 | Dated human fossil specimens from Mladeč. a, Upper jaw fragment Mladeč 8, male, exhibiting modern human features and some 'archaic' features such as dental dimensions. **b**, Cranium Mladeč 1, female, exhibiting derived modern human features. The graduation marks on both scales indicate centimetres. Copyright for the photographic material: Wolfgang Reichmann (2004), Naturhistorisches Museum, Anthropologische Abteilung, Burgring 7, 1010 Vienna, Austria.

[1]VERA (Vienna Environmental Research Accelerator) Laboratory, Institut für Isotopenforschung und Kernphysik der Universität Wien, Währingerstraße 17, [2]Department für Chemische Ökologie und Ökosystemwissenschaften, Universität Wien, Althanstrasse 14, A-1090 Wien, Austria. [3]Anthropologische Abteilung, Naturhistorisches Museum Wien, Burgring 7, A-1010 Wien, Austria. [4]Department of Anthropology, Washington University, Campus Box 1114, St Louis, Missouri 63130, USA.

© 2005 Nature Publishing Group

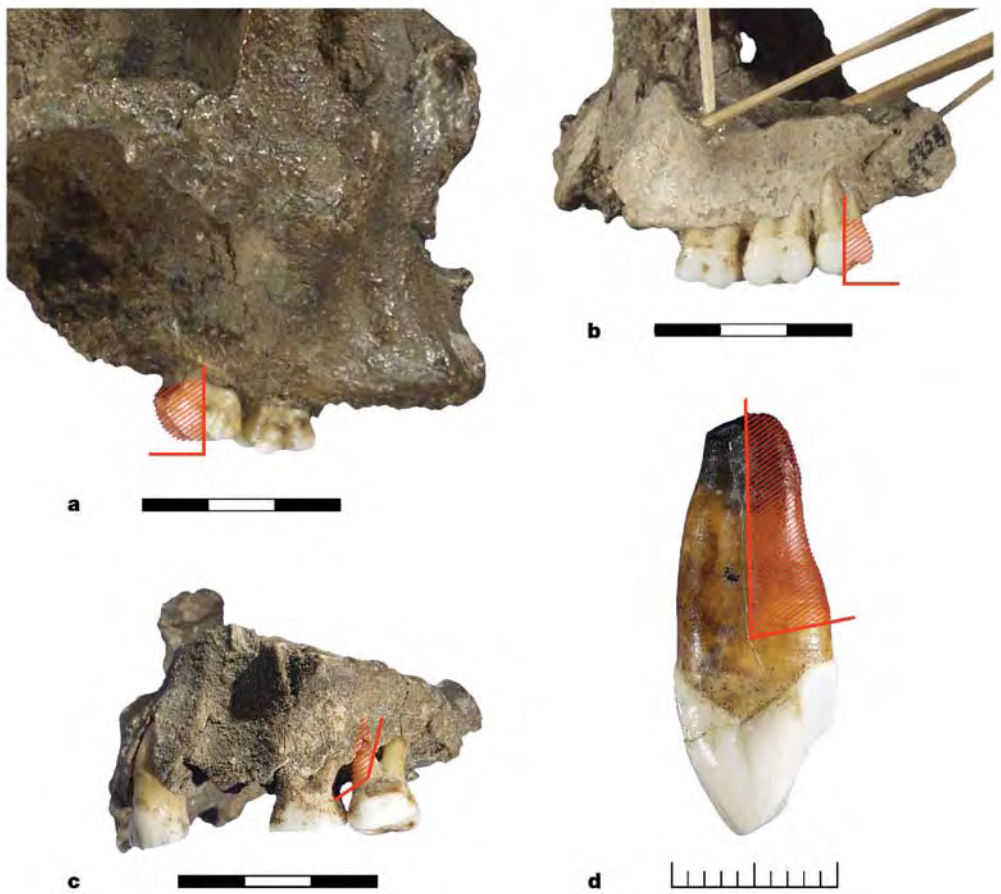


Figure 2 | Documentation of the dated specimens showing the sampled parts in red. **a**, Mladeč 1, lateral view from right. **b**, Mladeč 2, lateral view from left. **c**, Mladeč 8, lateral view from left. **d**, Mladeč 9a, right maxillary canine, mesial view. A centimetre scale is displayed for **a** to **c**; for **d** the minor graduation marks on the scale indicate 1 mm. Copyright for the photographic material: Wolfgang Reichmann (2004), Naturhistorisches Museum, Anthropologische Abteilung, Burgring 7, 1010 Vienna, Austria.

in the Mladeč 5 and 6 males, distinctive occipital bunning in Mladeč 3, 5 and 6, large palatal and dental dimensions of Mladeč 8 (Fig. 1a), the large crowns of the Mladeč 9a, 10 and 51 canines, and articular hypertrophy of some of the postcrania. Moreover, although they are robust compared to recent females, the Mladeč 1 (Fig. 1b) and 2 crania exhibit few of these features[5,6]. And most recently, ancient DNA from Mladeč 2 and 25c failed to show evidence of the mitochondrial (mt)DNA sequences found in some Neanderthals[18]. The Neanderthal affinities of some of the anatomical features have been questioned[7], but ultimately the implications of the Mladeč assemblage for human population dynamics with the transition to modern humans in Europe are dependent on whether they can be dated close to the transitional time period.

Several efforts have so far been made to obtain reliable and relevant ^{14}C dates from the Mladeč fossils, but all of them failed. These dating attempts were followed at the Centre for Isotope Research Radiocarbon Laboratory in Groningen by ^{14}C dating of carbonate samples from remnants of the crust that appear to have sealed the layer containing the human bones and artefacts in the 'Dome of the Dead'. From the results (GRN-26333: $34,160^{+520}_{-490}$ ^{14}C yr BP and GRN-26334: $34,930^{+520}_{-490}$ ^{14}C yr BP) it was concluded that the minimum age of the bones is 34–35 ^{14}C kyr BP[9]. Another attempt to date the human fossils was performed at the Vienna Environmental Research Accelerator (VERA) Laboratory in Vienna. Curatorial considerations to save the

Table 1 | Results of EA/IRMS quality checks of the Mladeč samples

Laboratory number	Sample name	C content (% DW)	N content (% DW)	C/N ratio	δ^{13}C (‰)	δ^{15}N (‰)
VERA-2736*	Mladeč 25c	6.44 ± 0.20	0.47 ± 0.06	13.7 ± 1.7	−24.6 ± 0.2	10.0 ± 0.5
VERA-3073†	Mladeč 1	11.8	3.2	3.7	−19.1	10.6
VERA-3074†	Mladeč 2	6.4	1.4	4.7	−20. 6	10.3
VERA-3075‡	Mladeč 8	10.7 ± 0.1	2.3 ± 0.2	4.7 ± 0.4	−21.4 ± 0.3	11.7 ± 0.4
VERA-3075‡	Mladeč 8, collagen	44.3 ± 0.3	16.1 ± 0.7	2.7 ± 0.1	−20.1 ± 0.4	10.9 ± 0.7
VERA-3076A‡ and VERA-3076B‡	Mladeč 9a, right maxillary canine	9.6 ± 0.6	2.4 ± 0.4	4.0 ± 0.3	−19.7 ± 0.2	9.6 ± 0.6

*Values represent the mean value of three EA/IRMS measurements and one standard deviation (s.d.) of the mean.
†Only a single EA/IRMS measurement was performed.
‡Values represent the mean value of two EA/IRMS measurements and one s.d. of the mean.
DW, dry weight. The δ^{13}C and δ^{15}N values are defined as the relative deviation (in ‰) of the ^{13}C/^{12}C and ^{15}N/^{14}N ratio of a sample from the ^{13}C/^{12}C of the V-PDB (Vienna-Pee Dee Belemnite) standard and the ^{15}N/^{14}N of the atmospheric N_2 standard, respectively. The s.d. of the mean values given in the table for multiple measured samples include uncertainties due to sample inhomogeneities. The reproducibility of the measurements of the laboratory standard was 0.1‰ (s.d.) for δ^{13}C and <0.2‰ (s.d.) for δ^{15}N. The δ^{13}C of the untreated samples reflects the isotopic composition of the total carbon present in the sample originating from the organic and the inorganic sample fraction, as well as from exogenous carbon. Similarly, the δ^{15}N reflects the isotopic composition of the total nitrogen.

© 2005 Nature Publishing Group

Table 2 | Radiocarbon ages determined for the human remains from the Mladeč site

Laboratory number	Sample name	Sample material	^{14}C-age (yr BP)*
VERA-2736	Mladeč 25c	Ulna	26,330 ± 170
VERA-3073	Mladeč 1	Right M2, distal half of the crown	$31,190^{+400}_{-390}$
VERA-3074	Mladeč 2	Left M3, distal half of the crown	$31,320^{+410}_{-390}$
VERA-3075	Mladeč 8	Left M2, mesial-buccal root	$30,680^{+380}_{-360}$
VERA-3076A	Mladeč 9a, right maxillary canine	Lingual half of the root (white-coloured collagen)	$31,500^{+420}_{-400}$
VERA-3076B	Mladeč 9a, right maxillary canine	Lingual half of the root (brown-coloured collagen)	27,370 ± 230

*Errors are one-sigma uncertainties.

human bones necessitated first ^{14}C dating of the animal remains from Mladeč with accelerator mass spectrometry (AMS), thereby dating the human remains indirectly. Out of eight selected samples from different species, only five could be dated successfully. The uncalibrated ^{14}C ages of these specimens yielded a wide range from 8.5 ^{14}C kyr BP to about 42 ^{14}C kyr BP (see Supplementary Table 1)—so an accurate indirect dating of the human remains was impossible. Therefore we needed to date the human remains directly.

We used a proximal ulna fragment, Mladeč 25c, and tooth samples from the most prominent specimens kept at the Naturhistorisches Museum in Vienna, that is, Mladeč 1, 2, 8 and the isolated maxillary right canine Mladeč 9a (samples from Mladeč 2 and Mladeč 25c were also used in the DNA study, see above). We assumed that the collagen in dentine is preserved from degradation and contamination (for example, consolidants) in non-abraded teeth with an intact enamel layer (Mladeč 1, 2) or in tooth roots covered by an intact alveolar bone (Mladeč 8). The isolated canine Mladeč 9a was in an excellent state of preservation in general and was therefore selected for dating as well. Before sampling, casts of the teeth were made to preserve the morphological information. Approximately one-half of each crown or part of the root (see Fig. 2) was taken for the radiocarbon determinations.

Amino-acid analysis of bone samples performed in the course of the DNA study indicated the variable preservation state of the human fossils. Amino acids of Mladeč 2 and 25c were identified as well preserved and fulfilled the criteria for DNA analysis, whereas the collagen of other bone samples, including Mladeč 8, appeared not suitable[18]. Although these results show that the collagen of some of the Mladeč fossils is well preserved, the suitability for ^{14}C dating of the selected samples was tested directly. A good preservation of the collagen was detected for all teeth samples via carbon and nitrogen elemental and isotopic analysis (see Methods and Table 1). Only the ulna appeared less well preserved and probably contaminated.

To remove possible superficial contaminants from the ^{14}C samples the tooth surface was abraded, leaving 350 to 200 mg of sample material for further processing. We used a method similar to that of ref. 19 for the chemical pre-treatment of the teeth (see Methods). The pretreated collagen from the human teeth and gelatine produced from the ulna and the animal bones were subjected to the routine sample preparation and measurement procedure used for ^{14}C dating of archaeological samples at VERA[20,21].

The ^{14}C ages of all human Mladeč samples dated in this study are listed in Table 2. All uncalibrated ages of the teeth agree at ~31 ^{14}C kyr BP within uncertainties (except for sample VERA-3076B). The ^{14}C age of ~26 ^{14}C kyr BP of the ulna is significantly younger. It is not certain that contamination in the ulna, detected in the quality check (see Methods), was completely removed by the applied chemistry. In the case of the isolated canine, a dark brown colour of the root apex was detected. Samples VERA-3076A and VERA-3076B both originate from the canine, but from different fractions of the extracted collagen—coloured white and brownish, respectively. The younger age of the dark collagen (VERA-3076B) supports our hypothesis that the colour of the apex resulted from contamination that was at least partially present after the chemical processing.

The ages determined for the Mladeč samples all lie within a time period for which a generally agreed calibration curve for the transformation of uncalibrated ^{14}C ages >20 kyr BP into calendar time ranges is not yet available. According to the existing, albeit divergent, ^{14}C records for this period determined in different archives, a shift of the 'true ages' by several thousand years towards higher ages might be possible[22]. (See refs 23 and 24 for further discussion on this issue.)

The AMS ^{14}C dating of the Mladeč human remains confirms that they derive from the time period of the middle to late Aurignacian of Central Europe. With the presence of multiple individuals, males and females, adult and immature with cranial, dental and post-cranial elements, the Mladeč assemblage is the oldest directly dated substantial assemblage of modern human remains in Europe. Only the ~35 ^{14}C kyr BP Peştera cu Oase mandible and cranium, from two individuals, are securely older among European early modern humans, and they currently lack postcranial remains and an archaeological association. Moreover, the Mladeč dates on both robust 'males' (Mladeč 8 and 9a) and less robust 'females' (Mladeč 1 and 2) fall into the same time period, reinforcing the idea that the variability within the assemblage reflects not only the original population variability but probably also its level of sexual dimorphism.

METHODS

Quality test of the sample material. In addition to the ^{14}C samples, small samples (~10 mg) of the dentine as well as some material from the ulna were acquired for combined elemental analysis/stable isotope ratio mass spectrometry (EA/IRMS). The measurements were performed with an elemental analyser (EA 1110, CE Instruments) coupled to a gas isotope ratio mass spectrometer (Delta^{PLUS}, Finnigan MAT) operating in the continuous flow mode.

We measured the carbon and nitrogen content and deduced the C/N ratios of the powdered samples. The C/N ratio comprises the ratio of the total carbon (from the inorganic matrix, the organic material and possible contaminants) to the total nitrogen (from amino acids and possible contaminants). This quality check was suggested by Hedges and van Klinken[25]. They argue that C/N ratios $\gg 4$ in bone samples may result from deamination (a diagenetic consequence) or may indicate the presence of large amounts of exogenous carbon. The C/N ratios of the untreated teeth lie between ~4.7 and ~3.7, and they are therefore within the accepted range or only slightly enhanced. This indicates that no significant carbon contamination is present in the samples.

The nitrogen content of fresh, defatted and dried compact bones lies between 4% and 5% DW (dry weight) (ref. 26). The good preservation state of the collagen of all teeth samples was denoted by N contents above 1.4% DW. Although mainly applied for collagen extracts and appraised as not very sensitive for the detection of contaminants, the δ^{13}C and δ^{15}N values of the untreated dentine support the good preservation of the collagen and the absence of large amounts of exogenous carbon (see Table 1).

The collagen yield of the ^{14}C sample Mladeč 8 (25 mg collagen extracted from ~300 mg sample material) was sufficiently high to enable an EA/IRMS measurement of a subsample from the purified collagen also. A C/N ratio of ~2.7 was determined for this sample. This value is comparable with C/N ranges from 3.5–2.9 (refs 27, 28) and 3.4–2.6 (ref. 29) for gelatinized collagen from well-preserved bones. The values of the N content and δ^{13}C and δ^{15}N values (see Table 1) also

© 2005 Nature Publishing Group

indicate that the collagen extracted from Mladeč 8 can be considered well preserved.

The low N content of the unprocessed ulna (~0.5% DW) and the low collagen yield during sample preparation show that the degradation of the collagen is already advanced, but the extant amount is still higher than 5% of the collagen content in recent bones. This 5% level is considered a prerequisite for reliable ^{14}C dating of bones[25]. A C/N ratio of about 14 indicates the presence of exogenous carbon in this sample.

Chemical pretreatment of ^{14}C teeth samples. We extracted the collagen from the teeth by dissolving the inorganic material in dilute HCl. In the case of the crown samples, where the major part of the enamel was removed mechanically, the rest of the enamel still present was dissolved in this step. (Collagen yields cannot be taken as a measure of the collagen preservation in the crowns because an unknown variable part of the enamel, which is very low in organic material, was present in these samples.) In addition to the demineralization step, we treated the samples with a dilute NaOH solution, and thereafter again with HCl. After each step the collagen was washed with bi-distilled water.

The gelatine production—frequently used for the preparation of ^{14}C-bone samples[26]—was omitted in the treatment of the human teeth to avoid unnecessary loss of the valuable sample material which comes along with each clean-up step. As mentioned above, for collagen originating from 'protected' dentine a smaller risk for contamination can be assumed. Even the NaOH treatment of the collagen for the removal of humic acids would not have been necessary, because the collagen appeared as a white coloured substance (except in one case, see above) after demineralization. The NaOH solution stayed uncoloured during this treatment, which indicates that no significant amount of humic acids was present in these samples.

Received 24 October 2004; accepted 22 March 2005.

1. Szombathy, J. Die diluvialen Menschenreste aus der Fürst-Johanns-Höhle bei Lautsch in Mähren. *Die Eiszeit* **2,** 1–34, 73–95 (1925).
2. Oliva, M. *UISPP XIIe Congres* **Vol. 2,** 207–216 (Institut d'Archéologie, Bratislava, 1993).
3. Conard, N. J., Grootes, P. M. & Smith, F. H. Unexpectedly recent dates for human remains from Vogelherd. *Nature* **430,** 198–200 (2004).
4. Svoboda, J. The depositional context of the Early Upper Paleolithic human fossils from the Koněprusy (Zlatý kůň) and Mladeč Caves, Czech Republic. *J. Hum. Evol.* **38,** 523–536 (2000).
5. Frayer, D. W. Cranial variation at Mladeč and the relationship between Mousterian and Upper Paleolithic hominids. *Anthropos* **23,** 243–256 (1986).
6. Wolpoff, M. H., Hawks, J., Fayer, D. W. & Hunley, K. Modern human ancestry at the peripheries: A test of the replacement theory. *Science* **291,** 293–297 (2001).
7. Bräuer, G., Collard, M. & Stringer, C. On the reliability of recent tests of the Out of Africa hypothesis for modern human origins. *Anat. Rec.* **279A,** 701–707 (2004).
8. Mellars, P. Neanderthals and the modern human colonization of Europe. *Nature* **432,** 461–465 (2004).
9. Svoboda, J. A., van der Plicht, J. & Kuzelka, V. Upper Palaeolithic and Mesolithic human fossils from Moravia and Bohemia (Czech Republic): some new ^{14}C dates. *Antiquity* **76,** 957–962 (2002).
10. Smith, F. H., Trinkaus, E., Pettitt, P. B., Karavanić, I. & Paunović, M. Direct radiocarbon dates for Vindija G_1 and Velika Pećina Late Pleistocene hominid remains. *Proc. Natl Acad. Sci. USA* **96,** 12281–12286 (1999).
11. Terberger, T., Street, M. & Bräuer, G. Der menschliche Schädelrest aus der Elbe bei Hahnöfersand und seine Bedeutung für die Steinzeit Norddeutschlands. *Archäol. Korrespond.* **31,** 521–526 (2001).
12. Henry-Gambier, D. Les fossils de Cro-Magnon (Les Eyzies-de-Tayac, Dordogne): nouvelles données sur leur position chronologique et leur attribution culturelle. *Bull. Mem. Soc. Anthropol. Paris* **14,** 89–112 (2002).
13. Orschiedt, J. Datation d'un vestige humain provenant de La Rochette (Saint Léon-sur-Vézère, Dordogne) par la méthode du carbone 14 en spectrométrie de masse. *Paléo* **14,** 239–240 (2002).
14. Trinkaus, E. *et al.* An early modern human from the Peştera cu Oase, Romania. *Proc. Natl Acad. Sci. USA* **100,** 11231–11236 (2003).
15. Stringer, C. B. In *Hominid Remains: An Update. British Isles and Eastern Germany* (ed. Orban, R.) 1–40 (Université Libre de Bruxelles, Brussels, 1990).
16. Păunescu, A. *Paleoliticul şi Mezoliticul din Spaţiul Transilvan* 231 (Editura AGIR, Bucharest, 2001).
17. Hublin, J. J., Spoor, F., Braun, M., Zonneveld, F. & Condemi, S. A late Neanderthal associated with Upper Palaeolithic artefacts. *Nature* **381,** 224–226 (1996).
18. Serre, D. *et al.* No evidence of Neanderthal mtDNA contribution to early modern humans. *PLoS Biol.* **2,** e57 (2004).
19. Schmitz, R. *et al.* The Neanderthal type site revisited: Interdisciplinary investigations of skeletal remains from the Neander Valley, Germany. *Proc. Natl Acad. Sci. USA* **99,** 13342–13347 (2002).
20. Wild, E. M. *et al.* First ^{14}C results from archaeological and forensic studies at the Vienna Environmental Research Accelerator. *Radiocarbon* **40,** 273–281 (1998).
21. Steier, P. *et al.* Pushing the precision limit of ^{14}C AMS. *Radiocarbon* **46,** 5–16 (2004).
22. Bard, E., Rostek, F. & Ménot-Combes, G. A better radiocarbon clock. *Science* **303,** 178–179 (2004).
23. van Andel, T. H., Davies, W., Weninger, B. & Jöris, B. In *Neanderthals and Modern Humans in the European Landscape During the Last Glaciation: Archaeological Results of the Stage 3 Project* (eds van Andel, T. H. & Davies, W.) 21–31 (Oxbow, Oxford, 2004).
24. van der Plicht, J. *et al.* NOTCAL04-Comparison/calibration ^{14}C records 26–50 cal kyr BP. *Radiocarbon* **46,** 1–14 (2004).
25. Hedges, R. E. M. & van Klinken, G. J. A review of current approaches in the pretreatment of bone for radiocarbon dating by AMS. *Radiocarbon* **34,** 279–291 (1992).
26. Petchey, F. in *Chemical Pretreatment Methods* (ed. Higham, T.) ⟨http://www.c14dating.com/pret.html⟩.
27. DeNiro, M. J. Postmortem preservation and alteration of *in vivo* bone collagen isotope ratios in relation to palaeodietary reconstruction. *Nature* **317,** 806–809 (1985).
28. Ambrose, S. H. Preparation and characterization of bone and tooth for isotopic analysis. *J. Archaeol. Sci.* **17,** 431–451 (1990).
29. Schoeninger, M. J., Moore, K. M., Murray, M. L. & Kingston, J. D. Detection of bone preservation in archaeological and fossil samples. *Appl. Geochem.* **4,** 281–292 (1989).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Frayer and G. Kurat for comments on an earlier version of this paper. We also thank H. Prossinger and other colleagues for discussions. Furthermore we thank W. Reichmann for the photographic documentation, R. Mühl and A. Walch for technical support, and S. Lehr for his help in sample preparation.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.M.W. (eva.maria.wild@univie.ac.at) and M.T.-N. (maria.teschler@univie.ac.at).

© 2005 Nature Publishing Group

LETTERS

Distinguishing random environmental fluctuations from ecological catastrophes for the North Pacific Ocean

Chih-hao Hsieh[1], Sarah M. Glaser[1], Andrew J. Lucas[1] & George Sugihara[1]

The prospect of rapid dynamic changes in the environment is a pressing concern that has profound management and public policy implications[1,2]. Worries over sudden climate change and irreversible changes in ecosystems are rooted in the potential that nonlinear systems have for complex and 'pathological' behaviours[1,2]. Nonlinear behaviours have been shown in model systems[3] and in some natural systems[1,4–8], but their occurrence in large-scale marine environments remains controversial[9,10]. Here we show that time series observations of key physical variables[11–14] for the North Pacific Ocean that seem to show these behaviours are not deterministically nonlinear, and are best described as linear stochastic. In contrast, we find that time series for biological variables[5,15–17] having similar properties exhibit a low-dimensional nonlinear signature. To our knowledge, this is the first direct test for nonlinearity in large-scale physical and biological data for the marine environment. These results address a continuing debate over the origin of rapid shifts in certain key marine observations as coming from essentially stochastic processes or from dominant nonlinear mechanisms[1,9,10,18–20]. Our measurements suggest that large-scale marine ecosystems are dynamically nonlinear, and as such have the capacity for dramatic change in response to stochastic fluctuations in basin-scale physical states.

Recent effort to characterize the decadal-scale behaviour of North Pacific physical and biological phenomena has centred on the concept of 'regime shifts'[18–21]. These regime shifts appear as quasi-stationary states in measured parameters, separated by periods of rapid transition[20]. Although attention has been focused on the qualitative phenomenology of these shifts (that is, documenting the appearance of distinct regimes with rapid shifts between them), nowhere in the discussion has their dynamical origin been directly assessed. True regime shifts are not random features of the time series, but are formally associated with the ideas of nonlinear amplification[7], alternative basins of attraction[20,22], multiple stable states[2], hysteresis and fold catastrophe[1,23], all of which require the underlying dynamics to be nonlinear in origin. For example, while it is quite clear that major changes occurred in the commonly measured North Pacific abiotic and biotic indices around 1976–77 (for example, patterns of sea surface temperature (SST), fisheries landings data, zooplankton abundance and community composition)[18,21], the nature of such changes remains elusive. Are such changes indicative of the operation of nonlinear dynamics or are they features of the data that might arise stochastically?

Some researchers (predominantly physical oceanographers) have suggested that apparent shifts observed in key physical variables are not singular (nonlinear transitional) events but instead represent normal statistical deviations[9,10]. Insofar as similar features in marine physical observations can be reproduced stochastically as random events[10], it is not necessary to invoke complicated nonlinear mechanisms. In contrast, others (predominantly biologists) have been inclined to see rapid environmental shifts as a fundamentally nonlinear phenomenon with important ecological implications[1,18,19]. They view the changes in populations and community structure that occur across putative regimes as being more than passive linear tracking of environmental variability. Rather, they see the rapid shifts in biological variables as an amplified response to environmental change pushing the system into different local basins of attraction or alternative states[7,8,20]. Fold catastrophes are a special case for achieving alternative states that raise the possibility of hysteresis or non-symmetrical reversibility of ecosystem states (for example, where population crashes are easier to attain than recoveries)[1]. Such instability and irreversibility have become a cautionary tale for environmental management and policy makers, bringing nonlinear phenomena to the fore.

A major weakness of the current debate is its focus on the statistical phenomenology of regime shifts. This approach examines the timing and magnitude of hypothesized shifts in the time series to see if they represent statistically distinct states separated by rapid transitions[20]. Identifying these plateaus and transitions usually involves subjectivity at some level that is difficult to overcome (for example, specifying the timing of shifts). Techniques that promise solutions for this require too much data for the biological time series involved[20], and many of the approaches assume the existence of only two states—a simple but arbitrary assertion. It seems curious that although the phenomenon being debated is a nonlinear one, nowhere in the methodological debate is the question of nonlinearity explicitly examined. Insofar as it is the nonlinear basis of the putative shifts that give them their meaning, it should be illuminating to directly measure the observations to determine if they are in fact consistent with the necessary hypothesis of nonlinearity.

Here we test a suite of key physical and biological time series observations for the North Pacific basin. Our aim is not to examine particular events to see if they satisfy the statistical description of a regime; rather, we look at complete time series to see if the variations contained in the whole data series were nonlinear in origin.

We examine these data using established methods from nonlinear time series analysis that involve state space reconstruction with lagged coordinate embeddings (Takens' theorem)[5,7,8,24]. To determine whether a time series reflects linear or nonlinear processes, we compare the out-of-sample forecast skill of a linear model versus an equivalent nonlinear model. This involves a two-step procedure: (1) we use simplex-projection[5] to identify the best embedding dimension (the number of independent variables required to model the process)[5] (Fig. 1), and (2) we use this embedding in the S-map procedure[7,8,25,26] to assess the nonlinearity of the time series (Fig. 2). The method of S-maps relies on fitting a series of models

[1]Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093-0202, USA.

© 2005 Nature Publishing Group

(from linear to nonlinear) where the degree of nonlinearity is controlled by a local weighting parameter θ. Improved out-of-sample forecast skill with increasingly nonlinear models (larger θ) indicates that the underlying dynamics were themselves nonlinear[25]. The forecast protocol, which involves a blind evaluation of forecast skill, is a rigorous standard that avoids model over-fitting or arbitrary fits to the data (see Methods and Supplementary Information).

We analyse the key time series commonly associated with the North Pacific regime shift debate that have sufficient length for the methods to apply. Physical measurements include indices that collapse North Pacific basin-wide phenomena into single time series (for example, via empirical orthogonal functions), and coastal SST time series (Table 1). The Southern Oscillation Index (SOI)[13] is widely used for tracking the state of the El Niño/Southern Oscillation, which is the leading source of North Pacific interannual climate variations. The North Pacific Index (NPI)[12] and the Pacific Decadal Oscillation Index (PDO)[11] track the leading patterns of North Pacific sea-level pressure and SST variability, respectively. We chose the three longest daily coastal SST time series in the eastern Pacific (Scripps Pier, Pacific Grove and Farallones Islands)[14]. These time series are highly correlated with basin scale indices while also reflecting strong local dynamics. This collection of data is broadly representative of the large-scale physical state of the North Pacific basin over the twentieth century (Table 1).

Biological data analysed include annual commercial landings for Pacific salmon and trout (1938–2000)[17], the weekly Scripps Pier diatom record (1920–39)[5] and the California Cooperative Oceanic Fisheries Investigation (CalCOFI) time series for copepods[16] and larval fishes[15] (Table 2). In order to generate time series of sufficient length, composite CalCOFI series were created and then analysed (see Methods). These data have been part of the regime-shift literature, and form a representative collection for this analysis.

All of the major physical indices in Table 1 possess characteristics consistent with high dimensional or stochastic processes (for example, Fig. 1c). This simple characterization is true from weekly to annual timescales. They are well modelled as linear autoregressive (AR) processes of high order, showing no significant forecast improvement as the S-maps are tuned towards nonlinear solutions (for example, Fig. 2c).

In contrast, all of the biological data in Table 2 (both simple and composites) consistently exhibit nonlinear signatures. These series were best rendered in relatively low embedding dimensions (for example, Fig. 1d), and without exception showed improvement in out-of-sample forecast skill as the S-maps were tuned towards increasingly nonlinear solutions (for example, Fig. 2d). Although not all $\Delta\rho$ (change in forecast skill measured as a difference in the correlation coefficients) in Table 2 are significant, all show improvement, which is significant for the biological ensemble (binomial probability <0.001).

The fundamentally different way that regime shifts in the North Pacific have been viewed by physical oceanographers and biological oceanographers coincides with the different character of their respective time series. The physical time series do not appear to arise from low dimensional nonlinear processes, and the irregular features that have been hypothesized to indicate nonlinear deterministic regimes are better characterized as stochastic. At first sight, they do not show the required nonlinearity to allow the interpretation that the shift-like features are more than random events.

It is perhaps not surprising that some of the physical indices appear to be linear-stochastic insofar as they are constructed from linear combinations of observations. As linear aggregates (that is, averages or linear orthogonal functions), these indices may mask possible nonlinearities in individual physical measurements. For example, although station-based barometric pressures in temperate latitudes are highly nonlinear, it has been shown that this nonlinearity can vanish in larger spatial aggregates[6]. The dynamics of measles in Great Britain also exhibit this behaviour: individual cities display highly nonlinear infection rates, but these deterministic nonlinear effects appear as noise when the individual cities are aggregated into a single time series for the country as a whole[24]. Thus, simply because the PDO, NPI and SOI do not show nonlinear characteristics does

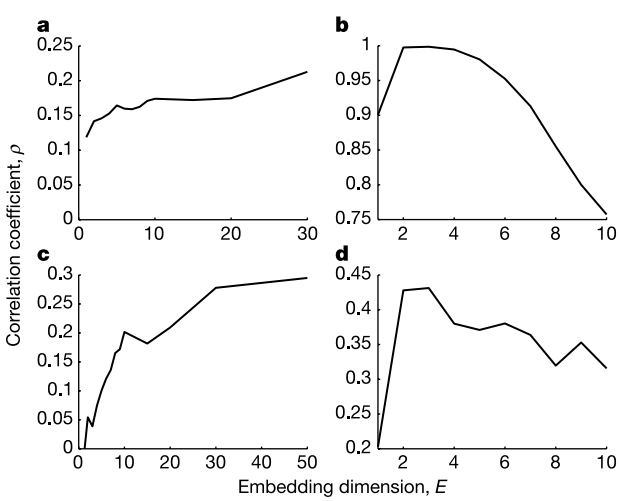


Figure 1 | Examples of the simplex projection method. a, b, Model time series; red noise[30] (**a**) and the nonlinear tent map[5] (**b**). **c, d,** Natural time series; Scripps Pier SST (**c**) and Scripps Pier diatoms (**d**). Panels **b** and **c** both show increasing skill (higher correlation coefficients, ρ) at higher embedding dimensions (E), which indicates that the underlying processes are high-dimensional (random for all practical purposes). In contrast, the chaotic tent map (**b**) and the Scripps Pier diatom time series (**d**) both show optimal skill (best ρ) with low-dimensional embeddings.

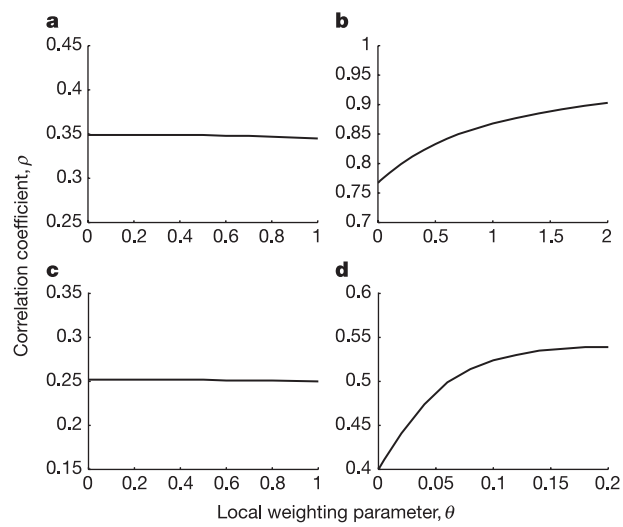


Figure 2 | Examples of the S-map method for the four time series from Fig. 1. The model generated red noise (**a**) and the Scripps Pier SST time series (**c**) show no improvement in forecast skill as S-maps are tuned towards increasingly local or nonlinear solutions (by increasing θ). Consequently, these time series do not show any indication of nonlinearity, and display all the hallmarks of a stochastic (high dimensional) linear generating mechanism. In contrast, the chaotic tent map (**b**) and the Scripps Pier diatom time series (**d**) both show increased skill as the S-map is tuned towards nonlinear solutions.

© 2005 Nature Publishing Group

Table 1 | Analyses of key North Pacific physical time series

Timescale	Physical data	Best E	Best θ	Best ρ	$\Delta\rho$	Nonlinear?	N	P-value
Weekly	SIO SST	20 +	0	0.252	0	No	4,226	1
Monthly	SIO SST	20 +	0	0.787	0	No	984	1
Monthly	Pacific Grove SST	20 +	0	0.524	0	No	945	1
Monthly	Farallones SST	20 +	0	0.486	0	No	764	1
Monthly	PDO	20 +	0	0.255	0	No	1,248	1
Monthly	NPI	20 +	0	0.636	0	No	1,260	1
Monthly	SOI	20 +	0	0.380	0	No	852	1
Quarterly	SIO SST	20 +	0	0.958	0	No	328	1
Quarterly	PDO	20 +	0	0.376	0	No	416	1
Quarterly	NPI	20 +	0	0.497	0	No	420	1
Quarterly	SOI	20 +	0	0.328	0	No	284	1
Annual	SIO SST, composite	20	0	0.770	0	No	984	1
Annual	PDO, composite	10	0	0.547	0	No	1,248	1
Annual	NPI, composite	16	0	0.674	0	No	1,260	1
Annual	SOI, composite	13	0	0.640	0	No	852	1

E, embedding dimension; θ, nonlinear tuning parameter. Best ρ (θ_{best}) indicates forecast skill (correlation coefficient), obtained using $\Delta\rho = ((\rho \text{ at } \theta_{\text{best}}) - (\rho \text{ at } \theta_0))$. A positive $\Delta\rho$ measures the difference in forecasting skill of the best nonlinear model (that is, where $\theta > 0$) as compared to the global linear model (that is, where $\theta = 0$). Thus, $\Delta\rho = \rho_{\text{best}} - \rho_0$. Data were analysed at various decimations (resolution in time scale). The PDO, NPI and SOI indices have monthly resolution. Quarterly data are averages of those monthly values. Daily coastal SST anomalies (daily data minus the year-day average over the entire record) were averaged to form weekly, monthly and quarterly time series. Owing to the paucity of data at the annual scale, we constructed composite time series by concatenating monthly values (all Januaries, all Februaries, ... all Decembers). These data are best embedded in high dimensions and show no improvement in forecast skill as the S-maps are tuned towards nonlinear solutions. As such, on timescales relevant to the regime shift debate, these physical oceanographic time series are unanimous in showing the hallmarks of linear stochastic generating mechanisms.

not preclude the possibility of nonlinear dynamics operating on finer scales. Nonetheless, these indices have been at the heart of the regime-shift debate, and we show their features to be stochastic.

More significantly, the various SST records, which are primary (non-aggregated) measurements, show the temperature shift phenomenon to be stochastic. That is, even simple SST measurements, which are emblematic of the physical regime shift phenomenon[21], do not indicate that temperature shifts have low dimensional nonlinear modes. Rather, they are high dimensional, and consequently they will be more difficult to model mechanistically in a way that replicates the phenomenological forecasting skill of a high degree AR model. This is a fundamental constraint on modelling efforts that we demonstrate empirically here. These findings resonate with the conception of the ocean as a linear red-noise integrator of atmospheric phenomena, a hypothesis first advanced in the 1970s[27]. However, it is clear that not all physical environmental time series are, by definition, high dimensional and stochastic; for example, analysis of meteorological observations shows strong low dimensional nonlinearity in the atmosphere, indicative of the potential for catastrophic climate change[6,28]. Furthermore, although true regime shift behaviour did not appear in the North Pacific physical oceanographic data that we tested, this obviously does not preclude the possibility of such behaviour having occurred further into the past or arising in the future. It simply did not occur in the last century, where the alleged shifts are indistinguishable from random events.

Biological time series appear to have dynamics that are fundamentally different from those of the physical variables associated with regime shifts. The relatively skilful out-of-sample forecasts at low embedding dimensions (even in composites) are consistent with the view that biological populations are nonlinear stochastic[25]. The full set of dynamics consists of a low dimensional, nonlinear, noise-free skeleton convolved with stochastic events acting on that skeleton to define the invariant measure[25]. Thus, the biological populations are not simply tracking the environment. Rather, our results support the hypothesis that ecological dynamics in the oceans can be characterized by nonlinear amplification of stochastic physical forcing by biological processes[7,8]. Regardless of interpretation, the biological time series for the North Pacific basin have the necessary signature for regimes to be actual nonlinear features of the data as opposed to randomly generated ones. This result for landings data and larval fish

Table 2 | Analyses of key North Pacific biological time series

Timescale	Biological data	Best E	Best θ	Best ρ	$\Delta\rho$	Nonlinear?	N	P-value
Weekly	Scripps Pier diatom	3	0.3	0.539	0.139*	Yes	830	<0.01
Monthly	Scripps Pier diatom	4	0.05	0.542	0.083	Yes	206	0.134
Quarterly	CalCOFI coastal larval fish	7	1.6	0.715	0.031*	Yes	3,220	<0.01
Quarterly	CalCOFI coastal-oceanic larval fish	8	0.6	0.744	0.017	Yes	1,400	0.164
Quarterly	CalCOFI oceanic larval fish	8	1.4	0.678	0.020*	Yes	4,760	0.040
Biannual	CalCOFI copepod	6	1.2	0.677	0.027	Yes	1,736	0.078
Annual	CalCOFI copepod	5	0.4	0.566	0.015	Yes	868	0.322
Annual	CalCOFI coastal larval fish	5	0.6	0.603	0.060*	Yes	805	0.038
Annual	CalCOFI coastal-oceanic larval fish	4	0.2	0.502	0.092	Yes	350	0.063
Annual	CalCOFI oceanic larval fish	7	0.6	0.576	0.017	Yes	1,190	0.273
Annual	Chinook salmon	3	0.4	0.448	0.440*	Yes	63	<0.01
Annual	Coho salmon	7	0.3	0.656	0.117	Yes	63	0.213
Annual	Chum salmon	4	0.18	0.634	0.767*	Yes	63	<0.01
Annual	Steelhead trout	3	0.2	0.281	0.272	Yes	63	0.118
Annual	Sockeye salmon	4	0.7	0.484	0.168	Yes	63	0.168
Annual	Composite salmon and trout	4	0.3	0.464	0.078	Yes	315	0.148

Parameters as in Table 1. Monthly diatom data are averages of weekly samples. Quarterly larval fish data represent four cruises per year, and biannual copepod data represent two cruises per year. Annual larval fish data are averages of the quarterly samples, and annual copepod data are averages of biannual samples. Commercial fisheries landing data are annual totals. These population data (described in text) are best embedded in low dimensions, and show improvement in forecast skill as the S-maps are tuned towards increasingly nonlinear solutions. Even where $\Delta\rho$ is not significant (asterisk indicates significant at the 0.05 level), the nonlinear model ($\theta > 0$) still outperforms the global linear model ($\theta = 0$). As such, these biological time series all show the hallmarks of nonlinear generating mechanisms.

© 2005 Nature Publishing Group

abundance should call into question static conceptions of maximum sustainable yield and the use of fixed exploitation quotas for managing commercial fisheries[2]. The potential for rapid and unpredictable shifts in response to environmental stochasticity and human impact supports a precautionary management approach for marine ecosystems[1].

METHODS

Forecasting techniques. To determine whether a time series reflects linear or nonlinear processes, we compare the out-of-sample forecast skill of a linear model versus an equivalent nonlinear model. This involves using lag coordinate embeddings in a two-step procedure: (1) we use simplex-projection[5] to identify the best embedding dimension, and (2) we use this embedding in the S-map procedure[7,8,25] to assess the nonlinearity of the time series. In both cases, model performance is evaluated out-of-sample with the time series divided into equal halves. The first half (library set) is used to build the model, while the second half (prediction set) is reserved to judge the out-of-sample performance of model forecasts.

Simplex projection is a nearest-neighbour forecasting algorithm that involves tracking the forward evolution of nearby points in a lag coordinate embedding[5]. For this study, an exploratory series of embedding dimensions (E) ranging from 1 to 20 (or higher) are used to evaluate the prediction, and the best E is chosen on the basis of prediction skill (Fig. 1). This embedding is then used in the S-map procedure. S-maps are an extension of standard linear autoregressive models in which the coefficients depend on the location of the predictee in an E-dimensional embedding[5,7,8,25,26]. (The predictee is the current state of the system from which the prediction is being made.) New coefficients are recalculated by singular value decomposition for each new prediction. In this calculation, the weight given to each vector in the library depends on how close that vector is to the predictee. The extent of this weighting is determined by the parameter θ. When $\theta = 0$ we obtain a global (single) linear map, and increasing values of θ in the S-map give increasingly local or nonlinear mappings[26] (Fig. 2). A detailed account of these methods is given in Supplementary Information.

All analyses were done both on raw values and on first differenced data to minimize the possibility of masking the nonlinear signal by trivial autocorrelation and to account for possible non-stationary trends in the data[5]. As no qualitative difference was found between analyses, we report here only the more conservative first differenced results.

Composite technique. The CalCOFI data represent one of the most comprehensive oceanographic monitoring programmes in the Pacific. Although hundreds of individual species are sampled, each time series alone is too short to apply the forecasting methods directly, particularly on the annual scale. For example, each larval fish time series contained only 35 annual data points (140 quarterly observations from 1951–2002, with a gap in quarterly collection and identification between 1967 and 1983). To accommodate the individual paucity of points due to these gaps, we generate composite time series[26] based on the known principal distributions of fish species[15,29]: coastal (23 taxa), coastal-oceanic (10 taxa), and oceanic groups (34 taxa). To minimize the number of gaps in the copepod data, only time series from 1951–66 and 1985–99 were used (for a total of 31 annual data points). Thus, we could only use copepod time series for the 28 taxa that occurred most frequently during the sampling period (at least 20 of 31 years). The copepods are treated as a single equivalence class, given the lack of an unambiguous separation of species by region. However, the results are unaffected when predominantly northern and predominantly southern species are treated provisionally as separate groups.

Individual time series are normalized to have unit mean and variance, and combined by equivalence class to produce composite time series[26]. This composite S-map procedure involves random combination of time series within each equivalence class (connecting individual time series end-to-end in different order to give different library/predictee combinations). The procedure is repeated 100 times or until all combinations are exhausted, and the average of these results for the CalCOFI data are reported in Table 2. The gaps and seams between time series are accounted for by discarding all vectors that traverse a gap or seam[26]. As a null test for the CalCOFI composites, we applied the procedure to null versions of each of the composite equivalence classes (composite time series with phases randomized) and as expected, obtained consistent linear signatures. By contrast, nonlinear signatures are obtained for the CalCOFI data even when the library and prediction halves are no longer randomly assigned, but are systematically chosen to be most different from each other (that is, with library and prediction sets each consisting of individual species whose time series covary most positively). This yields library sets of similar species that are most dissimilar to the prediction sets. In addition to the fact that all of the data in this study are normalized and first differenced, this additional test eliminates the remote possibility of producing a nonlinear artefact by combining heterogeneous data sets symmetrically into the library and prediction sets.

Physical variables required compositing at the annual scale only. The library vector comprises all January values from the first half of the time series (here ~1900–50) and so forth until December. The prediction vector is likewise constructed from the second half of the time series (here ~1951–2000).

Received 10 December 2004; accepted 14 March 2005.

1. Scheffer, M., Carpenter, S., Foley, J. A., Folkes, C. & Walker, B. Catastrophic shifts in ecosystems. *Nature* **413,** 591–596 (2001).
2. Steele, J. H. & Henderson, E. W. Modeling long-term fluctuations in fish stocks. *Science* **224,** 985–987 (1984).
3. May, R. M. Simple mathematical models with very complicated dynamics. *Nature* **261,** 459–467 (1976).
4. Ludwig, D., Jones, D. & Holling, C. S. Qualitative analysis of insect outbreak systems: the spruce budworm and the forest. *J. Anim. Ecol.* **47,** 315–332 (1978).
5. Sugihara, G. & May, R. M. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time-series. *Nature* **344,** 734–741 (1990).
6. Sugihara, G. *et al.* Residual delay maps unveil global patterns of atmospheric nonlinearity and produce improved local forecasts. *Proc. Natl Acad. Sci. USA* **96,** 14210–14215 (1999).
7. Dixon, P. A., Milicich, M. J. & Sugihara, G. Episodic fluctuations in larval supply. *Science* **283,** 1528–1530 (1999).
8. Dixon, P. A., Milicich, M. J. & Sugihara, G. in *Nonlinear Dynamics and Statistics* (ed. Mees, A. I.) 339–364 (Birkhauser, Boston, 2001).
9. Wunsch, C. The interpretation of short climate records, with comments on the North Atlantic and Southern Oscillations. *Bull. Am. Meteorol. Soc.* **80,** 245–256 (1999).
10. Rudnick, D. L. & Davis, R. E. Red noise and regime shifts. *Deep-Sea Res. I* **50,** 691–699 (2003).
11. Mantua, N. The Pacific Decadal Oscillation (PDO). ⟨http://jisao.washington.edu/pdo/PDO.latest⟩ (2004).
12. Hurrell, J. North Pacific (NP) Index information. ⟨http://www.cgd.ucar.edu/~jhurrell/np.html⟩ (2003).
13. National Center for Atmospheric Research. Southern Oscillation Index. ⟨http://www.cgd.ucar.edu/cas/catalog/climind/soi.html⟩ (2004).
14. SIO Shore Station. ⟨http://www.mlrg.ucsd.edu/shoresta/index.html⟩ (2001).
15. Hsieh, C. H. *et al.* A comparison of long-term trends and variability in populations of larvae of exploited and unexploited fishes in the Southern California region: a community approach. *Prog. Oceanogr.* (submitted).
16. Rebstock, G. A. Climatic regime shifts and decadal-scale variability in calanoid copepod populations off southern California. *Glob. Change Biol.* **8,** 71–89 (2002).
17. Washington Department of Fish and Wildlife and Oregon Department of Fish and Wildlife. *Status Report: Columbia River fish runs and fisheries 1938–2000* Table 3 ⟨http://www.dfw.state.or.us/ODFWhtml/InfoCntrFish/InterFish/2000statustables.pdf⟩ (2004).
18. Hare, S. R. & Mantua, N. J. Empirical evidence for North Pacific regime shifts in 1977 and 1989. *Prog. Oceanogr.* **47,** 103–145 (2000).
19. deYoung, B. *et al.* Detecting regime shifts in the ocean: data considerations. *Prog. Oceanogr.* **60,** 143–164 (2004).
20. Mantua, N. Methods for detecting regime shifts in large marine ecosystems: a review with approaches applied to North Pacific data. *Prog. Oceanogr.* **60,** 165–182 (2004).
21. McGowan, J. A., Bograd, S. J., Lynn, R. J. & Miller, A. J. The biological response to the 1977 regime shift in the California Current. *Deep-Sea Res. II* **50,** 2567–2582 (2003).
22. Sutherland, J. Multiple stable points in natural communities. *Am. Nat.* **108,** 859–873 (1974).
23. May, R. M. Biological populations with nonoverlapping generations: stable points, stable cycles, and chaos. *Science* **186,** 645–647 (1974).
24. Sugihara, G., Grenfell, B. & May, R. Distinguishing error from chaos in ecological time-series. *Phil. Trans. R. Soc. Lond. B* **330,** 235–251 (1990).
25. Sugihara, G. Nonlinear forecasting for the classification of natural time-series. *Phil. Trans. R. Soc. Lond. A* **348,** 477–495 (1994).
26. Sugihara, G., Allan, W., Sobel, D. & Allan, K. D. Nonlinear control of heart rate variability in human infants. *Proc. Natl Acad. Sci. USA* **93,** 2608–2613 (1996).
27. Hasselmann, K. Stochastic climate models. 1. Theory. *Tellus* **28,** 473–485 (1976).
28. Patil, D. A. S., Hunt, B. R. & Carton, J. A. Identifying low-dimensional nonlinear behavior in atmospheric data. *Mon. Weath. Rev.* **129,** 2116–2125 (2001).
29. Moser, H. G. *et al.* Distributional atlas of fish larvae and eggs in the Southern California Bight region: 1951–1998. *California Cooperative Oceanic Fisheries Investigations Atlas* **34,** 1–166 (2001).
30. Hastings, H. M. & Sugihara, G. *Fractals: A User's Guide for the Natural Sciences* (Oxford Univ., New York, 1993).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

© 2005 Nature Publishing Group

Acknowledgements We thank the Scripps Institution of Oceanography Pelagic Invertebrates Collection, the California Cooperative Oceanic Fisheries Investigation, the Oregon and Washington Departments of Fish and Wildlife, and G. Rebstock, for the availability and use of their data. R. Davis, D. Field, N. Mantua, G. Rebstock, C. Reiss, D. Rudnick, L.-F. Bersier and J. Bascompte provided discussion and comments on this work. Our study was funded by a NOAA initiative to improve the analysis of ecological data and its use in fisheries management (C.H.), the National Marine Fisheries Service and the Edna Bailey Sussman Fund (C.H.), NSF/LTER CCE 'Nonlinear Transitions in the California Current Coastal Pelagic Ecosystem' (C.H., G.S.), California Sea Grant (S.G.), the University of California Marine Council through the Network for Environmental Observations of the Coastal Ocean (A.L.), the Sugihara Family Trust (G.S.), and the Deutsche Bank Complexity Fund (G.S.). All co-authors contributed equally to this effort.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.S. (gsugihara@ucsd.edu).

© 2005 Nature Publishing Group

LETTERS

Sustained firing in auditory cortex evoked by preferred stimuli

Xiaoqin Wang[1], Thomas Lu[1], Ross K. Snider[1]† & Li Liang[1]

It has been well documented that neurons in the auditory cortex of anaesthetized animals generally display transient responses to acoustic stimulation, and typically respond to a brief stimulus with one or fewer action potentials[1–5]. The number of action potentials evoked by each stimulus usually does not increase with increasing stimulus duration[1,5–7]. Such observations have long puzzled researchers across disciplines and raised serious questions regarding the role of the auditory cortex in encoding ongoing acoustic signals. Contrary to these long-held views, here we show that single neurons in both primary (area A1) and lateral belt areas of the auditory cortex of awake marmoset monkeys (*Callithrix jacchus*) are capable of firing in a sustained manner over a prolonged period of time, especially when they are driven by their preferred stimuli. In contrast, responses become more transient or phasic when auditory cortex neurons respond to non-preferred stimuli. These findings suggest that when the auditory cortex is stimulated by a sound, a particular population of neurons fire maximally throughout the duration of the sound. Responses of other, less optimally driven neurons fade away quickly after stimulus onset. This results in a selective representation of the sound across both neuronal population and time.

The phenomenon that neurons in the auditory cortex under anaesthesia respond to sounds with phasic discharges irrespective of stimulus duration has been observed in a variety of mammalian species[1–7] with varying rates of discharge depending on the anaesthetic type and concentration[8–11]. The transient nature of auditory cortical responses and the lack of neural firing throughout stimulus duration have prompted researchers to propose various theories to explain neural coding strategies. For example, it has been suggested that neurons in the primary auditory cortex (A1) are specialized to respond to brief stimulus events[12]. A recent study suggested that A1 neurons discharge in a 'binary mode' in response to brief tone stimulation[5]. Researchers have also suggested that correlated firing between neurons, instead of the firing rates of individual neurons, signals the presence of steady-state sounds that fail to evoke continuing neural activity in A1 of anaesthetized marmosets[6] or cats[7]. On the other hand, accumulating evidence suggests that neurons recorded from the auditory cortex of awake animals exhibit both onset and sustained discharges in response to continuous acoustic stimulation[13–19]. However, the extent of sustained discharges in awake animals and their role in cortical coding of acoustic information remains unclear. Furthermore, as often encountered by experimenters, phasic responses are commonly observed even in the auditory cortex of awake animals.

In this report, we have addressed three specific questions regarding sustained firing in the auditory cortex of awake marmoset monkeys. First, can auditory cortex neurons discharge continuously over a prolonged period of time in response to continuous acoustic stimulation? Second, what are the stimulus conditions under which neurons in the auditory cortex are likely to discharge in a sustained manner? Third, do well-isolated single neurons in the auditory cortex discharge more than one action potential for each brief stimulus?

In contrast to observations from the auditory cortex of anaesthetized animals[6–7], we found neurons in both A1 and lateral belt (LB) areas of awake marmosets that were capable of firing throughout the duration of a tone or noise stimulus, over a prolonged period of time. Figure 1a shows an example of a single neuron responding to a pure tone at the neuron's best frequency. The tone was unmodulated and had a long duration (5 s) and slow rise and fall times (1 s). Such a stimulus generally produces the greatest adaptation in responses of auditory cortical neurons. In A1 of anaesthetized marmosets, this type of stimulus was shown to evoke only a brief response after stimulus onset[6]. In sharp contrast, the A1 neuron shown in Fig. 1a, which was recorded from an awake marmoset, gave strong, sustained discharges over the entire stimulus duration. As was typical of many A1 neurons we studied in awake marmosets, this single neuron was strongly driven by pure tones at its best frequency, but not by broadband noises. Neurons in the LB areas of marmosets, on the other hand, often responded more strongly to noise stimuli than to pure tones, similar to earlier observations in macaque monkeys[20]. Some LB neurons were found to respond continuously to unmodulated noise stimuli of long duration, as shown by the example in Fig. 1b to a 5-s broadband noise. Such prominent, sustained discharges have not been observed in the secondary auditory cortex of anaesthetized animals. In contrast to the A1 neuron shown in Fig. 1a, the LB neuron in Fig. 1b did not respond strongly to pure tones. Thus, the A1 and LB neurons shown discharged in a sustained manner when driven by their respective preferred stimulus (Fig. 1a, b).

The observations illustrated in Fig. 1 suggest that auditory cortex neurons can discharge in a sustained manner throughout stimulus duration if stimulated by their preferred stimuli. Pure tones and broadband noises were the extreme cases of a wide range of acoustic stimuli that could preferentially drive auditory cortex neurons in the awake condition. The majority of auditory cortex neurons we studied were preferentially driven by stimuli with intermediate spectral bandwidth and/or greater temporal complexity (for example, temporal modulations). We also studied neurons using stimuli with greater spectral and temporal complexity than pure tones and broadband noises. Most A1 neurons were more strongly driven by amplitude- or frequency-modulated tones than by pure tones, and typically showed preference for a particular, or 'best' modulation frequency (BMF). In contrast to anaesthetized animals, many A1 neurons in awake marmosets showed modulation frequency tuning without showing stimulus-synchronized discharges[19]. Similarly, neurons in LB areas often responded more strongly to amplitude-

[1]Laboratory of Auditory Neurophysiology, Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. †Present address: Department of Electrical and Computer Engineering, Montana State University, Bozeman, Montana 59717, USA.

© 2005 Nature Publishing Group

modulated noises at the BMF than to unmodulated noises. Neuron responses generally became weaker at modulation frequencies away from the BMF in both A1 and LB areas. Accompanying this reduction in the response magnitude was a change in temporal discharge pattern, from strong, sustained firing at the BMF to weakly sustained firing or onset firing as the modulation frequency moved away from the BMF. This trend was observed in both A1 (Fig. 2a) and LB areas (Fig. 2b). Note in Fig. 2 that although there is a large decrease in sustained firing between the preferred stimulus and non-preferred stimulus conditions, the change in onset firing between the two conditions is much smaller. In some neurons, off-responses emerged as the stimulus shifted from the preferred to non-preferred condition (Fig. 2b). Thus, when neurons in both cortical areas were driven by their preferred stimuli, they showed not only higher mean firing rates but also sustained firing patterns throughout the stimulus duration.

We further quantified the sustained and onset responses on a neuron-by-neuron basis (Fig. 3). We defined the first 100 ms of the response after the beginning of the stimulus as 'onset' (0–100 ms) and the remaining response as 'sustained' (from 100 ms until stimulus offset). Figure 3a compares the firing rates of the first and second halves of the sustained response at each neuron's BMF. The firing rates calculated from these two time intervals were similar, indicating that the sustained response of the neurons at their preferred stimuli indeed lasted throughout the stimulus duration. This property was observed for all three types of temporally modulated stimuli, regardless of whether or not neurons showed stimulus-synchronized firing patterns at the BMF (Fig. 3a). About 40% (45/113) of neurons showed discharges synchronized to the modulation envelope; the remaining ~60% of the tested neurons responded to temporally modulated stimuli at the BMF without showing stimulus-synchronized firing patterns. Note that such sustained firing differs from repetitive firing that results from entrainment to stimulus cycles; in the latter case, neurons would exhibit stimulus-synchronized firing patterns. We suggest that the

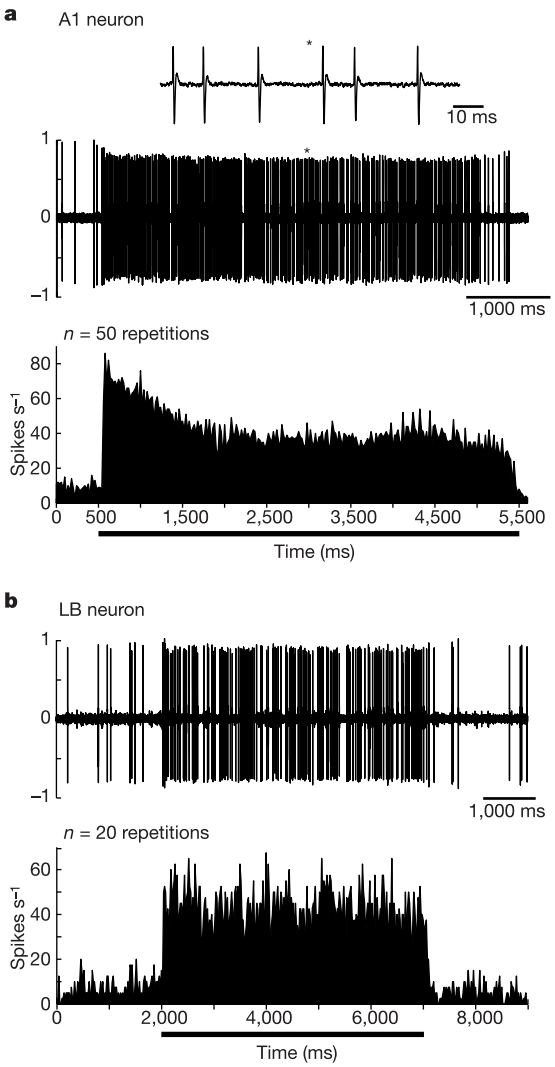


Figure 1 | Examples of sustained firing of single neurons evoked by long-duration stimuli. a, A1 neuron (unit M42M-171). Stimulus: unmodulated pure tone at the neuron's best frequency (9.3 kHz), 80 dB SPL, rise/fall time 1 s, duration 5 s. Upper panel, raw recording trace of the response to the 25th presentation of the stimulus. Inset shows an expanded view of a portion of the trace (2950–3050 ms). Asterisks mark the time at 3,000 ms. Lower panel, PSTH computed from responses to 50 repetitions of the same stimulus (bin-width, 20 ms). **b**, LB neuron (unit M60107-124). Stimulus: unmodulated broadband noise, 80 dB SPL, rise/fall time 5 ms, duration 5 s. Upper panel, response to 20th presentation of the stimulus. Lower panel, PSTH (20 repetitions; bin-width, 20 ms). The thick bar below the *x* axis in **a** and **b** (lower panels) indicates stimulus duration.

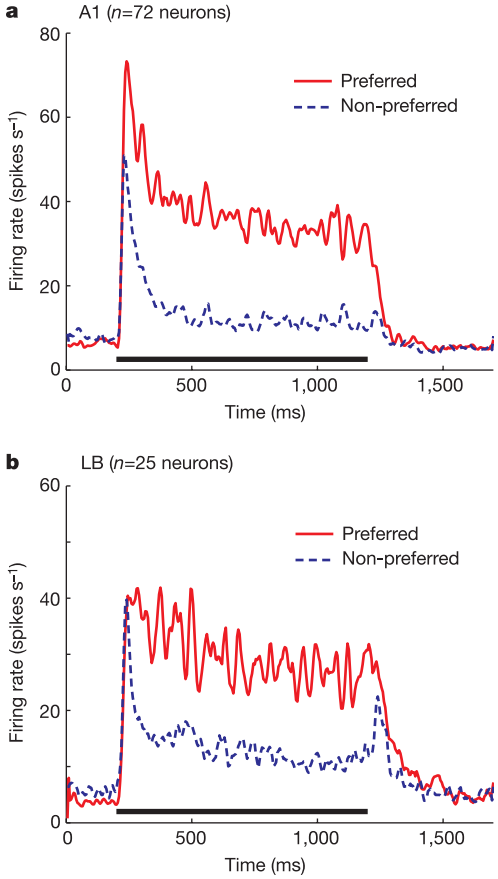


Figure 2 | Temporal firing patterns evoked by preferred and non-preferred stimuli. a, b, Mean PSTHs calculated from populations of A1 (**a**) and LB (**b**) neurons in response to each neuron's preferred stimulus (sAM, sFM or nAM) at preferred (red) and non-preferred (blue) modulation frequencies (10–20 repetitions at each modulation frequency). The preferred modulation frequency is equivalent to the best modulation frequency (BMF). The non-preferred modulation frequency corresponds to the minimum firing rate above the BMF. A similar trend is observed at the non-preferred modulation frequency below the BMF. PSTH bin-width, 5 ms (smoothed by a 5-point moving triangular window). The thick bar above the *x* axis indicates stimulus duration (1 s).

© 2005 Nature Publishing Group

sustained firing observed in these conditions truly represents transformed representations of time-varying acoustic signals. Additional examples of sustained responses to 5-s tone or noise stimuli were analysed and are shown in Fig. 3a.

As shown by averaged population post-stimulus time histograms (PSTHs) in Fig. 2, the magnitude of onset response relative to sustained response was larger at the non-preferred than at the preferred stimulus condition. We analysed this property quantitatively on a neuron-by-neuron basis (Fig. 3b) using an onset response index, defined as the ratio of onset versus sustained firing rates. This index is much higher when measured at the worst modulation frequency compared to the best modulation frequency (Fig. 3b), indicating a greater extent of onset responses when neurons are less optimally driven.

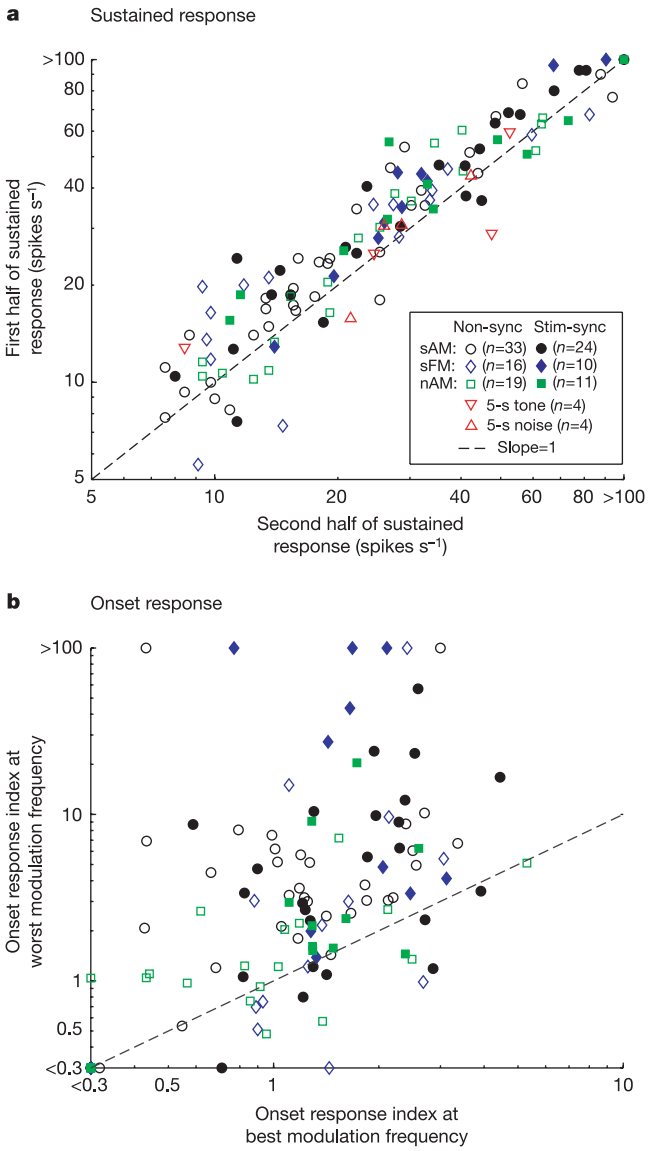


Figure 3 | Quantitative analysis of sustained and onset responses. Data include the 72 A1 neurons and 25 LB neurons analysed in Fig. 2, and an additional 16 neurons from the rostral and medial areas surrounding A1 (113 neurons in total), all tested with 1-s long, temporally modulated stimuli (sAM, sFM or nAM). Neurons showing stimulus-synchronized or non-synchronized discharges at BMF (quantified by the Rayleigh statistic) are marked with filled or open symbols, respectively. **a**, Mean firing rates of the first half (100–550 ms) and second half (550–1,000 ms) of the sustained response at each neuron's BMF are compared. Rates given as mean ± s.d. in spikes s^{-1} for the first half, second half: sAM, 36.3 ± 27.2, 30.7 ± 24.0 ($n = 57$); sFM, 35.7 ± 25.0, 29.9 ± 22.0 ($n = 26$); nAM, 37.3 ± 25.1, 33.5 ± 25.9 ($n = 30$). Sustained responses of 8 additional neurons to 5-s long, unmodulated best-frequency-tones or noises were also analysed (windows for first half: 100–2,550 ms; second half: 2,550–5,000 ms). **b**, The onset response index, the ratio of onset firing rate (0–100 ms) divided by the sustained firing rate (100–1,000 ms) is compared between the BMF and the worst modulation frequency for an individual neuron (corresponding to the minimum firing rate across all tested modulation frequencies above BMF). A larger index value indicates a greater onset response relative to the sustained response. Symbols used in **b** are defined in the key in **a**.

The complex stimuli used in the experiments described in Figs 2 and 3 had relatively long duration (1 s). Most previous studies of the auditory cortex have used brief, pure tone stimuli. We carried out additional experiments in awake marmosets to determine whether cortical responses evoked by short duration tones share the same or different properties with the responses recorded from anaesthetized animals. A recent study[5] using anaesthetized rats suggested that A1 neurons discharge in a 'binary mode', and the appearance of multiple action potentials per stimulus reported in previous studies could have resulted from multi-unit recording methods. The examples shown in Fig. 1 clearly demonstrate that prolonged, sustained firing can be observed in well-isolated single neurons from extracellular recordings in the awake condition. The majority of A1 neurons we studied did not discharge in a binary fashion in response to brief tones at a neuron's best frequency. Figure 4a shows the distribution of the number of action potentials per stimulus presentation for 50-ms best-frequency-tone stimulation. The majority of neurons discharged more than one action potential per stimulus presentation (median 2 per trial), and the distribution was well fit by a Poisson distribution. Furthermore, the number of action potentials per stimulus presentation generally increased with increasing stimulus duration. Figure 4b shows that A1 neurons responded to 100-ms best-frequency-tones with more action potentials per stimulus presentation (median 3 per trial) than to 50-ms best-frequency-tones. Thus, unlike single neurons in A1 of anaesthetized animals, which typically fire one or fewer action potential per stimulus irrespective of stimulus duration[1–5], single neurons in A1 of awake marmosets discharge in a Poisson-like manner when stimulated by brief (50 ms) best-frequency-tones. To further analyse discharge patterns evoked by tone stimuli, we compared the count of initial action potentials (from stimulus onset time to 10 ms after response latency) with that of all action potentials (from stimulus onset time to offset time) of tone-evoked discharges on a neuron-by-neuron basis (Fig. 4c). If a neuron had only initial action potentials, it should fall along the dashed line (slope of 1). If a neuron responded to each stimulus presentation with an average of one or fewer action potentials, it should fall within the shaded square in Fig. 4c. Note that nearly all neurons were located above the dashed line, indicating that they had sustained discharges beyond the initial action potentials. The median value for initial action potentials was 0.82 per trial (close to the value of ~1 per trial typically recorded in A1 of anaesthetized animals). Of the 263 neurons tested, 231 neurons (88%) had more than one action potential per stimulus. These data clearly show that A1 neurons in the awake condition do not discharge in a binary mode in response to brief tones. The distribution shown in Fig. 4a differs qualitatively from the similar measure obtained in anaesthetized rats[5].

The functional role of sustained discharges has been extensively studied in visual cortex[21], but much less in auditory cortex. Our findings show that whether a neuron responds to a stimulus with sustained discharges depends crucially on the optimality of the stimulus. Specifically, we suggest that auditory cortex neurons fire in a sustained manner when they are driven by their preferred stimuli, but show either mostly onset discharges or no discharges at all when stimulated by non-preferred stimuli. We found that neurons in the auditory cortex of awake marmosets are often highly selective to acoustic stimuli, and as such, the preferred stimulus (or stimuli) of a neuron only occupies a small region of

© 2005 Nature Publishing Group

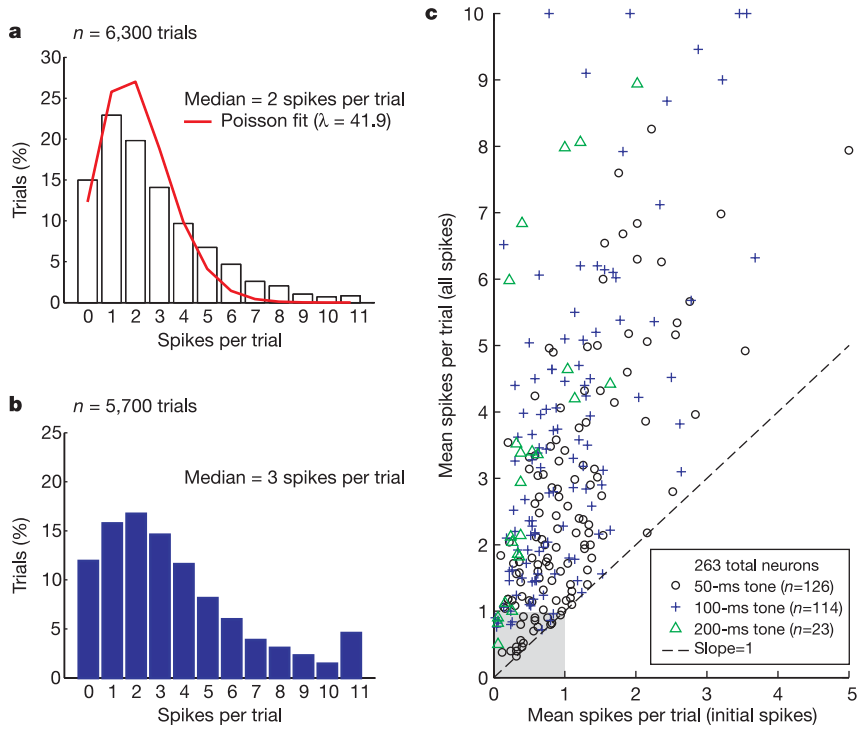


Figure 4 | Quantitative measures of single-neuron responses to brief tones. The tones were set at each neuron's best frequency and preferred sound level (for non-monotonic neurons) or 30 dB above threshold (for monotonic neurons), and delivered for 50 repetitions. **a**, Distribution of spikes per trial for 50-ms best-frequency-tone responses, calculated over the entire stimulus duration for 6,300 stimulus presentations obtained from 126 neurons (median 2 spikes per trial, equivalent to 40 spikes s^{-1}). The thick red curve represents the least-squares fit by a Poisson distribution ($\lambda = 41.9$ spikes s^{-1}, $T = 0.05$ s, $r = 0.97$). **b**, Same as **a** but for 100-ms best-frequency-tone responses obtained from 114 neurons (median 3 spikes per trial, equivalent to 30 spikes s^{-1}). **c**, Mean spikes per trial (per neuron), calculated over the entire stimulus duration ('all spikes') versus that calculated over the initial response period ('initial spikes', time window from 0 to response latency + 10 ms). The response latency was determined for each neuron using an accumulative spike count histogram.

the acoustic parameter space (defined by spectral, temporal and intensity axes). This explains why it is common for experimenters to encounter onset (phasic) responses in the auditory cortex even under unanaesthetized conditions.

The observations provided in this report have important implications for neural coding mechanisms in the auditory cortex. First, sustained discharges in the auditory cortex show greater stimulus specificity than onset discharges. Second, the lack of stimulus-synchronized firing patterns in the sustained responses indicates that such responses are transformed (instead of preserved) representations of time-varying acoustic signals. Third, the different temporal discharge patterns evoked by preferred and non-preferred stimuli suggest that auditory cortex neurons adapt to non-preferred stimuli more quickly than to preferred stimuli. These findings show that the auditory cortex is capable of dynamically responding to the acoustic environment with sustained firing from optimally driven neurons and onset firing from others. When a sound is heard, the auditory cortex first responds with transient (onset) discharges across a relatively large population of neurons. As time passes, the activation becomes restricted to a smaller population of neurons that are preferentially driven by the sound.

These observations have immediate implications for interpreting spatial and temporal activation patterns of the auditory cortex recorded by other techniques commonly applied to alert human subjects, for example, electroencephalography (EEG), positron emission tomography (PET) and functional magnetic resonance imaging (fMRI). The sustained discharges in the auditory cortex probably result from or are enhanced by intracortical processing, both within a cortical area (via recurrent or interlaminar connections) and from higher cortical areas (via efferent feedback connections). The underlying cellular and network mechanisms remain to be identified.

In contrast to sustained discharges, onset discharges in the auditory cortex might have different roles in representing the external acoustic environment. For example, widespread onset responses with short latency might allow the auditory cortex to detect quickly (albeit less discriminatively) the occurrence of sounds from prey, predators and the environment. The precision of the onset action potential timing is preserved throughout the ascending auditory pathway, from the auditory nerve[22] to the auditory cortex[12], and can thus provide precise timing information to mark the onset of a sound or transitions in an ongoing acoustic stream[15,23,24]. There could also be species-specific differences regarding the roles of onset and sustained discharges in cortical coding. For example, the presence of sustained discharges appears to be less prominent in bat auditory cortex even under unanaesthetized conditions[25]. The sonar signals of bats are extremely brief, and their communication sounds are much shorter[26] than species-specific vocalizations of marmosets[27] and other non-human primates. However, the extent of any sustained discharges would not be fully revealed if only brief stimuli were used in experiments.

METHODS

Animal preparation and recording procedures. Animal preparation and recording procedures have been detailed in previous publications from our laboratory[15,19]. Single-neuron activity was recorded from three awake marmosets trained to sit quietly during recording sessions, using tungsten microelectrodes (impedance 2–5 MΩ at 1-kHz), and continuously monitored by the experimenter while data recordings progressed. The signal-to-noise ratio of action potential waveforms was typically >10:1 in our recordings (often much greater). Judging by the depth of recording and response characteristics,

© 2005 Nature Publishing Group

the majority of recorded single neurons were from layers II–III and, to a lesser extent, from layer IV. The location of the A1 was determined by its tonotopic organization[28,29] and its relationship to the lateral belt area. Acoustic stimuli were generated digitally (sampling rate 100 kHz) and delivered in free-field through a loudspeaker located ~1 m in front of the animal, in a double-walled, soundproof chamber with the interior covered by 3 inches of acoustic absorption foam. Stimuli included pure tones, band-pass or broadband noises, sinusoidal amplitude- or frequency-modulated tones (sAM or sFM), and sinusoidal amplitude-modulated band-pass or broadband noises (nAM).

Determination of a neuron's stimulus preference. Single neurons recorded from the auditory cortex of awake marmosets are highly selective for stimulus parameters (in time, frequency and intensity domains). A neuron may require not only a proper carrier frequency, sound level and spectral bandwidth, but also a proper temporal modulation frequency in order to be maximally driven. We use the term 'preferred stimulus' to refer to stimuli that were locally optimal (evoking maximum firing rate) along one or more stimulus dimensions. For example, the preferred stimulus for an A1 neuron could be an sAM tone with a carrier frequency of 8 kHz, modulation frequency of 32 Hz and 60 dB sound pressure level (SPL). Because of technical difficulties involving sampling over multiple stimulus dimensions, the true optimal stimulus of a cortical neuron could not be easily determined using current analytical methodologies. The preferred stimuli used in our experiments were optimized in carrier frequency, sound level, temporal modulation frequency and spectral bandwidth, and could be considered the 'best stimulus' in this context, probably approaching the optimal stimulus of a neuron.

Once a single neuron was isolated, its preferred stimulus was determined using the following procedure. (1) We first determined a neuron's preference for tone or noise. A1 neurons were typically easily driven by tones, whereas LB neurons were more responsive to noises than tones, consistent with earlier observations in macaque monkeys[20]. (2) For tone-preferring neurons, a best frequency was determined using pure tones delivered at various frequencies (20–40 steps per octave, depending upon the neuron's sharpness of tuning). For noise-preferring neurons, a centre frequency was determined using band-pass noises at various centre frequencies, with a preferred bandwidth that maximized evoked responses. Stimuli were delivered at 10 dB above a neuron's estimated threshold. The best frequency estimated by tones was generally similar to the centre frequency estimated by band-pass noises in neurons that did not show a clear preference for tone or noise. (3) The rate-level function of a neuron at its best frequency or centre frequency was then obtained (0–100 dB SPL in 5–10-dB steps) using the stimulus optimized in the above procedures. Similar to previous observations in awake macaque monkeys[30], a large proportion (~70%) of single neurons in A1 or LB in awake marmosets had non-monotonic rate-level functions for which a preferred sound level could be determined. For neurons with monotonic rate-level functions, the threshold sound level at best frequency or centre frequency was determined. (4) The preferred temporal modulation frequency was determined using sAM (or sFM, depending upon a neuron's preference) or nAM stimuli, for tone- or noise-preferring neurons, respectively, at various modulation frequencies (typically 4–512 Hz, in octave steps). The carrier frequency was set at a neuron's best frequency (or centre frequency). The sound level was set at the preferred level (for non-monotonic neurons) or 30 dB above threshold (for monotonic neurons). For sAM and nAM stimuli, the modulation depth was 75–100%. For sFM stimuli, the frequency modulation depth was set at a preferred value for the neuron. For nAM stimuli, a preferred bandwidth was chosen for the noise carrier. A majority of A1 and LB neurons showed preference for a particular modulation frequency. Some neurons did not respond to unmodulated tones or noises and could only be driven by sAM, sFM or nAM stimuli. For these neurons, their best frequency (or centre frequency) was determined by varying the carrier frequency of temporally modulated stimuli. When necessary, the above process was iterated to optimize all stimulus parameters.

Data analysis. For neurons tested with temporally modulated stimuli (Figs 2 and 3), a total of 158 single neurons recorded from three marmosets showed band-pass tuning by mean firing rate to modulation frequencies above 10 Hz. Sustained responses at BMF to at least one of the test stimuli were observed in 113 neurons (71.52%). A sustained response was defined as having a minimum firing rate of 5 spikes s^{-1} during both the first and second halves of the stimulus (1-s stimulus duration). Out of 113 neurons, 72 neurons were recorded from A1, 25 from LB and 16 from the rostral and medial areas surrounding A1. The median spontaneous firing rates were 2.98 spikes s^{-1} (A1 neurons), 3.83 spikes s^{-1} (LB neurons) and 3.1 spikes s^{-1} (all neurons). The distribution of spontaneous firing rates of all 113 neurons can be fitted by an exponential function $y = 20.53\exp(-0.27x)$, $(r = 0.92)$.

A total of 263 single neurons were studied with short duration best-frequency-tones in A1 of three animals (Fig. 4). Responses of all neurons were significantly above spontaneous firing ($P < 0.01$, Wilcoxon rank sum test). The median spontaneous firing rate of this population of neurons was 1.8 spikes s^{-1} (25–75% percentile range, 0.7–4.68 spikes s^{-1}). The distribution of spontaneous firing rates of all 263 neurons can be fitted by the exponential function $y = 57.66\exp(-0.28x)$, $(r = 0.97)$.

Received 21 December 2004; accepted 18 March 2005.

1. Phillips, D. P. Temporal response features of cat auditory cortex neurons contributing to sensitivity to tones delivered in the presence of continuous noise. *Hear. Res.* **19**, 253–268 (1985).
2. Calford, M. B. & Semple, M. N. Monaural inhibition in cat auditory cortex. *J. Neurophysiol.* **73**, 1876–1891 (1995).
3. Heil, P. Auditory cortical onset responses revisited. II. Response strength. *J. Neurophysiol.* **77**, 2642–2660 (1997).
4. Schnupp, J. W., Mrsic-Flogel, T. D. & King, A. J. Linear processing of spatial cues in primary auditory cortex. *Nature* **414**, 200–204 (2001).
5. DeWeese, M. R., Wehr, M. & Zador, A. M. Binary spiking in auditory cortex. *J. Neurosci.* **23**, 7940–7949 (2003).
6. deCharms, R. C. & Merzenich, M. M. Primary cortical representation of sounds by the coordination of action-potential timing. *Nature* **381**, 610–613 (1996).
7. Eggermont, J. J. Firing rate and firing synchrony distinguish dynamic from steady state sound. *Neuroreport* **8**, 2709–2713 (1997).
8. Zurita, P., Villa, A. E., de Ribaupierre, Y., de Ribaupierre, F. & Rouiller, E. M. Changes of single unit activity in the cat's auditory thalamus and cortex associated to different anesthetic conditions. *Neurosci. Res.* **19**, 303–316 (1994).
9. Gaese, B. H. & Ostwald, J. Anesthesia changes frequency tuning of neurons in the rat primary auditory cortex. *J. Neurophysiol.* **86**, 1062–1066 (2001).
10. Cheung, S. W., Nagarajan, S. S., Bedenbaugh, P. H., Schreiner, C. E., Wang, X. & Wong, A. Auditory cortical neuron response differences under isoflurane versus pentobarbital anesthesia. *Hear. Res.* **156**, 115–127 (2001).
11. Bar-Yosef, O., Rotman, Y. & Nelken, I. Responses of neurons in cat primary auditory cortex to bird chirps: effects of temporal and spectral context. *J. Neurosci.* **22**, 8619–8632 (2002).
12. Phillips, D. P. Neural representation of stimulus times in the primary auditory cortex. *Ann. NY Acad. Sci.* **682**, 104–118 (1993).
13. Bieser, A. & Müller-Preuss, P. Auditory responsive cortex in the squirrel monkey: neural responses to amplitude-modulated sounds. *Exp. Brain Res.* **108**, 273–284 (1996).
14. Recanzone, G. H. Response profiles of auditory cortical neurons to tones and noise in behaving macaque monkeys. *Hear. Res.* **150**, 104–118 (2000).
15. Lu, T., Liang, L. & Wang, X. Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nature Neurosci.* **4**, 1131–1138 (2001).
16. Mickey, B. J. & Middlebrooks, J. C. Representation of auditory space by cortical neurons in awake cats. *J. Neurosci.* **23**, 8649–8663 (2003).
17. Malone, B. J., Scott, B. H. & Semple, M. N. Context-dependent adaptive coding of interaural phase disparity in the auditory cortex of awake macaques. *J. Neurosci.* **22**, 4625–4638 (2002).
18. Brugge, J. F. & Merzenich, M. M. Responses of neurons in auditory cortex of the macaque monkey to monaural and binaural stimulation. *J. Neurophysiol.* **36**, 1138–1158 (1973).
19. Liang, L., Lu, T. & Wang, X. Neural representations of sinusoidal amplitude and frequency modulations in the primary auditory cortex of awake primates. *J. Neurophysiol.* **87**, 2237–2261 (2002).
20. Rauschecker, J. P., Tian, B. & Hauser, M. Processing of complex sounds in the macaque nonprimary auditory cortex. *Science* **268**, 111–114 (1995).
21. Shadlen, M. N. & Newsome, W. T. Noise, neural codes and cortical organization. *Curr. Opin. Neurobiol.* **4**, 569–579 (1994).
22. Rhode, W. S. & Smith, P. H. Encoding timing and intensity in the ventral cochlear nucleus of the cat. *J. Neurophysiol.* **56**, 261–286 (1986).
23. Elhilali, M., Fritz, J. B., Klein, D. J., Simon, J. Z. & Shamma, S. A. Dynamics of precise spike timing in primary auditory cortex. *J. Neurosci.* **24**, 1159–1172 (2004).
24. Lu, T. & Wang, X. Information content of auditory cortical responses to time-varying acoustic stimuli. *J. Neurophysiol.* **91**, 301–313 (2004).
25. Suga, N. Principles of auditory information-processing derived from neuroethology. *J. Exp. Biol.* **146**, 277–286 (1989).
26. Kanwal, J. S., Matsumura, S., Ohlemiller, K. & Suga, N. Analysis of acoustic elements and syntax in communication sounds emitted by mustached bats. *J. Acoust. Soc. Am.* **96**, 1229–1254 (1994).
27. Wang, X. On cortical coding of vocal communication sounds in primates. *Proc. Natl Acad. Sci. USA* **97**, 11843–11849 (2000).
28. Aitkin, L. M., Merzenich, M. M., Irvine, D. R., Clarey, J. C. & Nelson, J. E. Frequency representation in auditory cortex of the common marmoset (*Callithrix jacchus jacchus*). *J. Comp. Neurol.* **252**, 175–185 (1986).
29. Wang, X., Merzenich, M. M., Beitel, R. & Schreiner, C. E. Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: temporal and spectral characteristics. *J. Neurophysiol.* **74**, 2685–2706 (1995).

© 2005 Nature Publishing Group

30. Pfingst, B. E. & O'Connor, T. A. Characteristics of neurons in auditory cortex of monkeys performing a simple auditory task. *J. Neurophysiol.* **45,** 16–34 (1981).

Acknowledgements We thank A. Pistorio for assistance with animal training and care, S. Eliades for technical assistance, R. Adams for help with maintaining the marmoset colony, and C. Wen for support throughout this work. This work was supported by an NIH grant and a US Presidential Early Career Award for Scientists and Engineers (X.W.).

Author Contributions X.W. led the development of the chronic recording preparation for awake marmosets with contributions from co-authors. R.K.S. and T.L. co-designed the Xblaster stimulus delivery and data acquisition system used in this study. X.W., L.L. and T.L. carried out the electrophysiological recording experiments. X.W. and T.L. analysed the data and wrote the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to X.W. (xwang@bme.jhu.edu).

© 2005 Nature Publishing Group

LETTERS

Regulation of PDGF signalling and vascular remodelling by peroxiredoxin II

Min Hee Choi[1]*, In Kyung Lee[1]*, Gyung Whan Kim[2], Bang Ul Kim[1], Ying-Hao Han[3], Dae-Yeul Yu[3], Hye Sun Park[1], Kyung Yong Kim[4], Jong Seo Lee[4], Chulhee Choi[1,6], Yun Soo Bae[1], Byung In Lee[2], Sue Goo Rhee[5] & Sang Won Kang[1]

Platelet-derived growth factor (PDGF) is a potent mitogenic and migratory factor that regulates the tyrosine phosphorylation of a variety of signalling proteins via intracellular production of H_2O_2 (refs 1–3). Mammalian 2-Cys peroxiredoxin type II (Prx II; gene symbol *Prdx2*) is a cellular peroxidase that eliminates endogenous H_2O_2 produced in response to growth factors such as PDGF and epidermal growth factor[4]; however, its involvement in growth factor signalling is largely unknown. Here we show that Prx II is a negative regulator of PDGF signalling. Prx II deficiency results in increased production of H_2O_2, enhanced activation of PDGF receptor (PDGFR) and phospholipase Cγ1, and subsequently increased cell proliferation and migration in response to PDGF. These responses are suppressed by expression of wild-type Prx II, but not an inactive mutant. Notably, Prx II is recruited to PDGFR upon PDGF stimulation, and suppresses protein tyrosine phosphatase inactivation. Prx II also leads to the suppression of PDGFR activation in primary culture and a murine restenosis model, including PDGF-dependent neointimal thickening of vascular smooth muscle cells. These results demonstrate a localized role for endogenous H_2O_2 in PDGF signalling, and indicate a biological function of Prx II in cardiovascular disease.

Members of the 2-Cys peroxiredoxins—a novel family of peroxidases that reduce hydroperoxides with the use of reducing equivalents provided by thioredoxin—are found in various cellular compartments[5,6]. Prx II has a high affinity for H_2O_2 (Michaelis constant (K_m) for H_2O_2 $<10\,\mu M$)[7] and is abundant in cytoplasm, making it a prime candidate as a regulator of the H_2O_2 signal initiated by growth factors. Upon PDGF stimulation, Prx $II^{-/-}$ mouse embryonic fibroblasts (MEFs) exhibited approximately twice the level of H_2O_2 production at the 10-min peak compared with wild-type cells, as measured by 2′,7′-dichlorofluorescein fluorescence[4] (Fig. 1a). H_2O_2 then returned to basal levels within 30 min. Total PDGF-induced tyrosine phosphorylation was markedly elevated in Prx $II^{-/-}$ MEFs compared with wild-type cells (Fig. 1b), coinciding with the time course and amplitude of H_2O_2 production. We next analysed the activation of downstream signalling molecules by using phospho-specific antibodies. Phospholipase Cγ1 (PLC-γ1) phosphorylation at Y783 and Y1253, residues known to be essential for activation[8], was elevated in Prx $II^{-/-}$ MEFs compared to wild-type cells (for pY783 blot: 9.7 ± 0.1 versus 3.9 ± 0.4 at 5 min; 8.2 ± 0.1 versus 2.8 ± 0.2 at 10 min, $P < 0.0001$, where units are fold increases ± s.d. in pY783 versus unstimulated control). However, activation-dependent phosphorylation of c-Src, ERK and Akt were unchanged (Fig. 1c). Two other clonal pairs of wild type and Prx $II^{-/-}$ MEFs gave the same result (data not shown), demonstrating reproducibility between the MEF clones. Similar results were also observed in epidermal growth factor-treated MEFs (Supplementary Fig. 1a). An 'add-back' rescue experiment was then carried out by retroviral expression of wild-type human Prx II in Prx $II^{-/-}$ MEFs. The retroviral-induced expression level of Prx II in Prx $II^{-/-}$ MEFs was similar to that of endogenous Prx II in wild-type MEFs (Fig. 2d). Re-introducing Prx II via retroviral expression appreciably suppressed protein tyrosine phosphorylation, including that of PLC-γ1, and abolished PDGF-induced inositol-1,4,5-trisphosphate generation (Fig. 1d, e). Furthermore, retroviral expression of Prx II reduced PDGF-induced proliferation as well as the chemotactic migration and adhesion induced by PDGF (Fig. 1f; see also Supplementary Fig. 1b). Collectively, these results indicate that Prx II is a novel component in PDGF signalling.

In addition to Prx II, mammalian cells also have two other peroxidase types capable of reducing H_2O_2 to water: catalase and glutathione peroxidase. We therefore investigated the influence of these peroxidases on PDGF signalling by using specific inhibitors. Treatment with 3-amino-1,2,4-triazole (ATZ), a specific catalase inhibitor, and buthionine-S,R-sulphoximine (BSO), a glutathione-depleting reagent, reduced the activity of endogenous catalase by 95% at 20 mM and the glutathione level by 90% at 50 μM, respectively (data not shown). Although both reagents resulted in an approximately twofold increase in the level of H_2O_2 (Supplementary Fig. 2), the level of PDGF-induced tyrosine phosphorylation and PLC-γ1 activation was not increased (Fig. 1g). These results suggest that Prx II is the main, if not the sole, and specific cellular regulator of PDGF signalling.

Next, in order to understand the molecular mechanism by which PDGF signalling is upregulated in Prx $II^{-/-}$ MEFs, we analysed the activation of PDGFR. We produced phospho-specific PDGFR antibodies against the peptides encompassing the seven known tyrosine phosphorylation sites (each of which provides a docking site for SH2 domain-containing signalling molecules such as c-Src, the p85 subunit of phosphatidylinositol-3-OH kinase, GAP, Grb2, SHP-2 and PLC-γ) on the receptor (ref. 1 and Supplementary Fig. 3). Notably, in contrast to wild-type cells, Prx $II^{-/-}$ MEFs displayed markedly elevated phosphorylation of PDGFR residues Y579/581 and Y857 (Fig. 2a). Immunoblot analyses using commercial anti-phospho-PDGFR-β antibodies also reproduced the above results (Supplementary Fig. 4). Immunoprecipitation experiments further confirmed that site-selective hyper-phosphorylation had occurred on PDGFR-β (Fig. 2b). Retroviral expression of wild-type Prx II in Prx

[1]Division of Molecular Life Sciences and the Center for Cell Signaling Research, Ewha Womans University, Seoul 120-750, Korea. [2]Department of Neurology, Yonsei University College of Medicine, Seoul 120-752, Korea. [3]Laboratory of Human Genomics, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-333, Korea. [4]Life Science Institute, Labfrontier Co. Ltd., KSBC building, san 111-8, lui-dong, Paldal-gu, Suwon, Kyunggi-do 442-766, Korea. [5]Laboratory of Cell Signaling, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [6]Department of Biosystems, KAIST, Daejeon 305-701, Korea.
*These authors contributed equally to this work.

© 2005 Nature Publishing Group

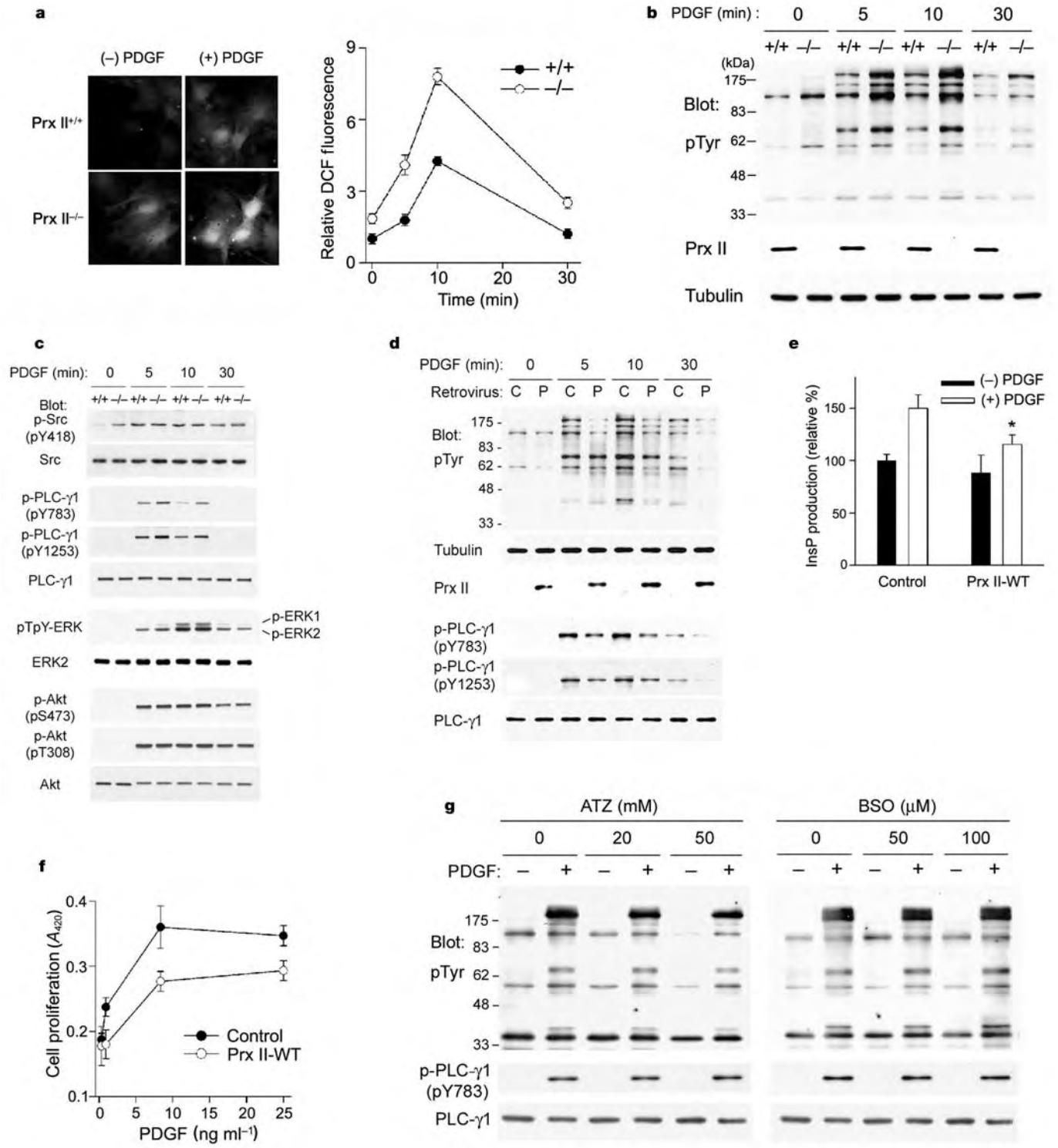


Figure 1 | Prx II is a physiological negative regulator of PDGF signalling. **a**, 2′,7′-Dichlorofluorescein (DCF) fluorescence image at 10 min stimulation. Quantitative data are means ± standard deviation (s.d.) of values from five random fields and representative of two independent experiments. **b**, Immunoblot analysis of protein tyrosine phosphorylation with anti-phosphotyrosine antibody (4G10, Upstate). **c**, Phosphorylation profiles of signalling molecules downstream of PDGFR-β in MEFs. **d**, Tyrosine phosphorylation of whole protein and PLC-γ1 in Prx II$^{-/-}$ MEFs infected with control (C) or Prx II wild-type retrovirus (P). **e**, PDGF-induced inositol phosphate (InsP) production in Prx II$^{-/-}$ MEFs infected with control or Prx II wild-type retrovirus. Data are represented as a per cent increase ± s.d. of the quantity of InsP versus unstimulated control ($n = 3$; asterisk, $P < 0.02$ versus stimulated control). **f**, PDGF-induced proliferation in Prx II$^{-/-}$ MEFs infected with control or Prx II wild-type retrovirus. Data are mean optical densities at 420 nm (A_{420}) ±s.d. ($n = 4$). **g**, Effect of catalase inhibition and glutathione depletion on PDGF-induced tyrosine phosphorylation. In **b**–**d** and **g**, representative immunoblots are shown ($n = 3$).

© 2005 Nature Publishing Group

II$^{-/-}$ MEFs reduced the PDGFR phosphorylation at Y579/581 and Y857 (Fig. 2c). In contrast, the Prx II inactive mutant, in which two active-site cysteine residues (C51 and C172) were substituted to serine residues, did not (Fig. 2d). This result indicates that the peroxidase activity of Prx II is essential for regulation of PDGFR phosphorylation. Because phosphorylation of both Y579/581 and Y857 sites are crucial for full activation of PDGFR kinase[9], the *in vitro* kinase activity of PDGFR-β was measured using glutathione *S*-transferase (GST)–PLC-γ1 as the direct substrate. Prx II$^{-/-}$ MEFs exhibited a marked PDGF-dependent increase in receptor kinase activity towards PLC-γ1 (Fig. 2e). Together, these results suggest that accumulation of the phosphorylated PLC-γ1 in Prx II$^{-/-}$ MEFs is the result of hyper-activation of PDGFR kinase by cooperative action of Y579/581 and Y857, which leads to expeditious phosphorylation and release of PLC-γ1 bound to the Y1021 site[10,11]. In addition, we checked that c-Src was not involved in PDGFR activation in Prx II$^{-/-}$ MEFs by using a Src kinase inhibitor (PP2). H_2O_2 induces the ligand-independent activation of c-Src in some cell types[12,13], and the phosphorylated Y579/581 residues are known to be a docking site for c-Src. However, PP2 neither repressed the hyper-phosphorylation of PDGFR nor its autophosphorylation in Prx II$^{-/-}$ MEFs (Fig. 2f).

We next investigated whether the novel pattern of PDGFR activation resulting from Prx II deficiency is mimicked by exogenously added H_2O_2. To answer this, wild-type MEFs pre-treated with low doses of PDGF (5 ng ml^{-1}) were challenged with increasing

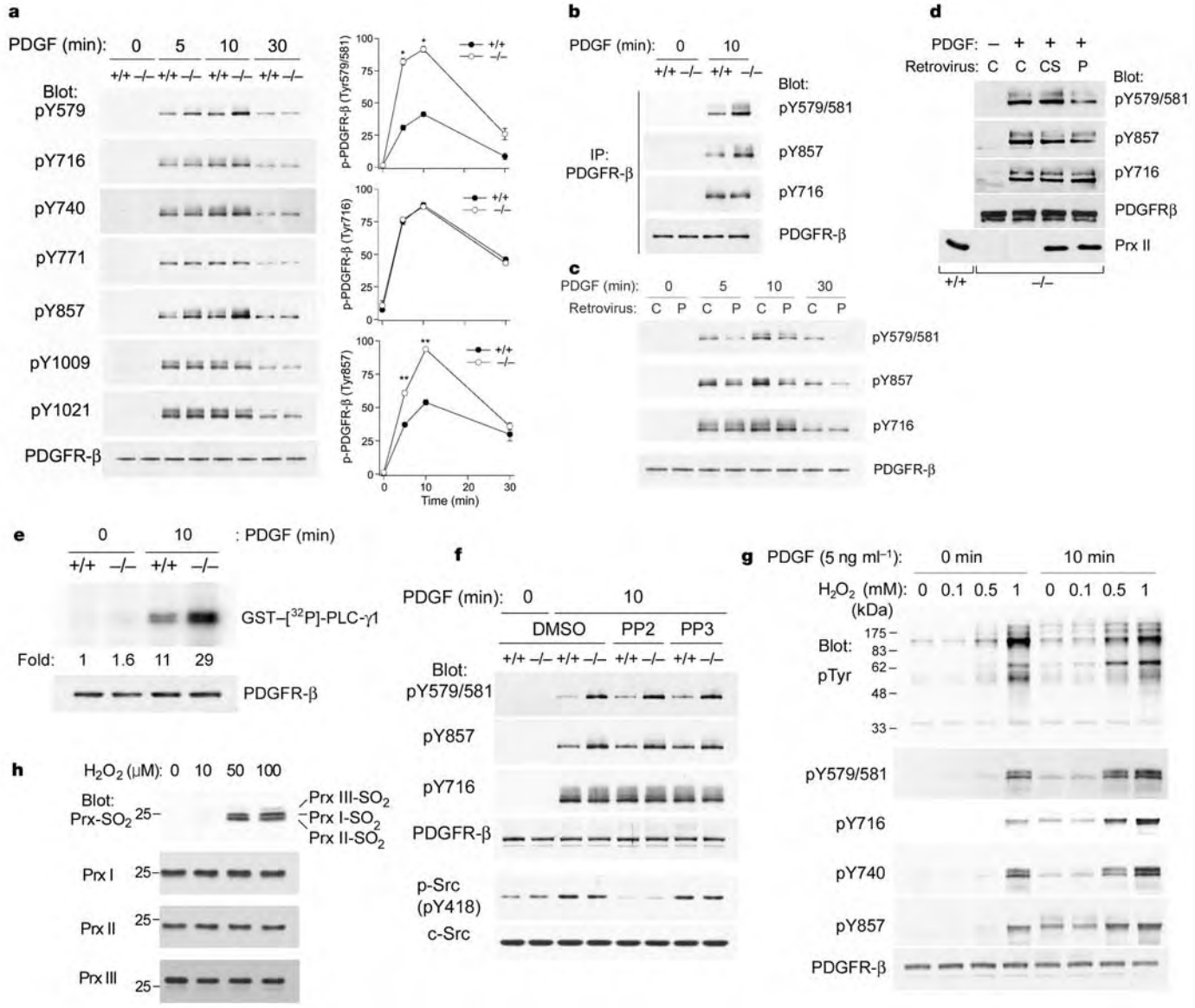


Figure 2 | Phosphorylation-site-selective regulation of PDGFR activation by Prx II. **a**, Immunoblot analysis of PDGFR-β phosphorylation using anti-phospho-PDGFR-β antibodies against the indicated tyrosine residues. Representative immunoblots are shown ($n = 4$). Graphs show quantified data (means ± s.d.; asterisk, $P < 0.002$, double asterisk, $P < 0.0001$ versus wild-type MEFs) for the amount of phosphorylated PDGFR-β after being normalized by the amount of non-phosphorylated PDGFR-β. **b**, Tyrosine phosphorylation of immunoprecipitated PDGFR-β. **c**, Tyrosine phosphorylation of PDGFR-β in Prx II$^{-/-}$ MEFs infected with control (C) or Prx II wild-type retrovirus (P). **d**, Effect of the Prx II cysteine mutant (CS) on PDGFR phosphorylation. The endogenous level of Prx II in wild-type MEFs is shown. **e**, *In vitro* kinase activity of PDGFR-β. Fold increase values are means of two independent experiments. **f**, Effect of a c-Src inhibitor (2 μM each, Calbiochem) on PDGFR-β phosphorylation. Note that c-Src activity was completely abolished by 30-min pre-treatment, as assessed by the extent of autophosphorylation at the Tyr 418 residue of c-Src. **g**, Effect of exogenous H_2O_2 on the tyrosine phosphorylation of pre-activated PDGFR-β. **h**, Analysis of 2-Cys peroxiredoxin over-oxidation by using the 2-Cys Prx sulphinic-acid-specific antibody[23]. In **g** and **h**, wild-type MEFs were treated with the indicated concentration of H_2O_2 for 10 min.

© 2005 Nature Publishing Group

concentrations of exogenous H_2O_2. The phosphorylation of PDGFR at the seven different tyrosine residues examined increased in response to addition of H_2O_2 in a concentration-dependent manner (Fig. 2g and data not shown). Moreover, exogenous H_2O_2 resulted in over-oxidation of 2-Cys peroxiredoxins, which indicates a loss of peroxidase activity[14], at a concentration as low as 50 μM (Fig. 2h). By contrast, such over-oxidation of 2-Cys peroxiredoxins was undetectable in both NIH3T3 and MEF cells treated with increasing concentrations of PDGF (data not shown). In addition, when the Prx II-overexpressing MEFs were treated with exogenous H_2O_2, overall phosphorylation of PDGFR was suppressed compared with that in normal MEFs (Supplementary Fig. 5), indicating that Prx II may scavenge H_2O_2 diffused into the cells. More importantly, this result is in contrast with the case of PDGF stimulation in NIH3T3 and smooth muscle cells overexpressing Prx II (see below). Together, these results suggest that exogenous H_2O_2 does not have a signal messenger role that endogenous H_2O_2 has. Furthermore, we could exclude the possibility that endogenous H_2O_2 influences the activity of unoccupied receptors in the same or adjacent cells.

We investigated further the functional role of Prx II in endogenous H_2O_2 signalling. Because H_2O_2 is small and diffusible, we hypothesized that Prx II and a target molecule of endogenous H_2O_2 work together spatially with the receptor to bring about the localized specificity. In order to demonstrate this, we first examined whether or not Prx II interacts with PDGFR. Co-immunoprecipitation experiments showed that, upon PDGF stimulation, Prx II is recruited to PDGFR-β in both MEFs and NIH3T3 cells, whereas Prx I is not (Fig. 3a). The interaction between Prx II and activated PDGFR-β gradually decreased after a 10-min peak, and eventually disappeared after 1 h (Fig. 3b). Furthermore, Prx II over-oxidation was not detected in the PDGFR immunoprecipitates (data not shown), indicating that the associated Prx II is in its active form. Immunostaining experiments also demonstrated PDGF-dependent co-localization of Prx II with PDGFR (Fig. 3c). These results suggest that Prx II can restrict the action radius of endogenous H_2O_2 close to the receptor. Because the receptor kinase itself is unlikely to discriminate between each site in a H_2O_2-dependent manner, a conceivable model might be the oxidative inactivation of protein tyrosine phosphatases (PTPases; the activities of which are reversibly regulated by stimulation-dependent oxidation[15]) near the receptor by endogenous H_2O_2. Therefore, we also examined whether Prx II influences PTPase oxidation in response to PDGF. Primarily, the membranous fraction (where PDGFR is exclusively present) and the soluble fraction were separated from NIH3T3 cells. As expected, some Prx II was also detected in the membranous fraction (Fig. 3d). We also asked whether Prx II overexpressed in NIH3T3 cells suppresses PDGFR phosphorylation in response to PDGF, as seen in Prx $II^{-/-}$ MEFs (Supplementary Fig. 6). The activities of oxidized PTPases (after recovery by dithiothreitol (DTT) reduction) in both the soluble and membranous fractions from NIH3T3 cells either overexpressing Prx II or not were measured using poly-(Glu_4-pTyr) peptides as the substrate. Notably, PTPase oxidation by PDGF occurred mainly in the membranous fraction and was significantly reduced by Prx II expression (Fig. 3e). This was confirmed in the membranous fraction from Prx $II^{-/-}$ MEFs retrovirally expressing wild-type Prx II (Fig. 3f). Taken together, the results suggest that Prx II translocated to PDGFR may relieve the membrane-associated PTPases from oxidative inactivation by eliminating H_2O_2 in the microenvironment of PDGFR.

PDGF is a pivotal factor for the proliferation and migration of

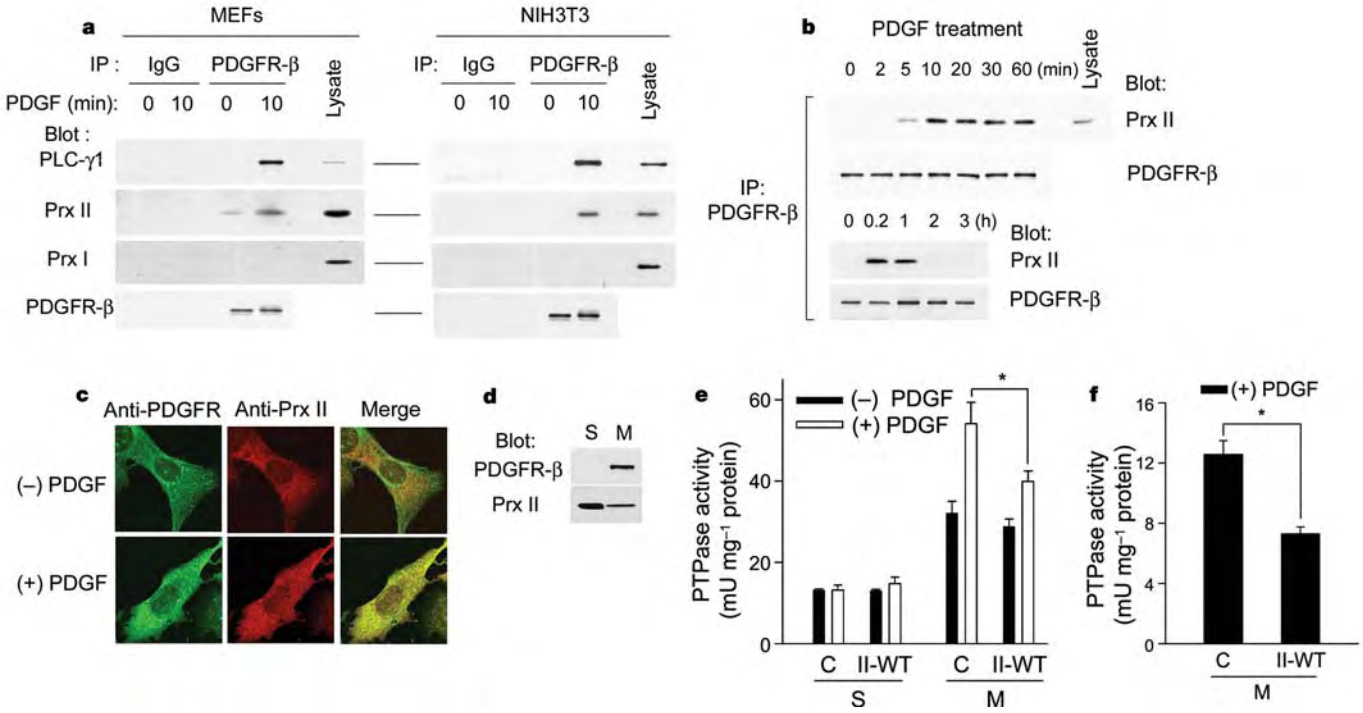


Figure 3 | Interaction of Prx II with activated PDGFR-β. **a**, Stimulation-dependent association of Prx II with PDGFR-β. **b**, Transient interaction of Prx II with PDGFR-β in NIH3T3 cells stimulated with PDGF. The immunoblots were obtained and normalized to the quantity of PDGFR immunoprecipitated. Representative immunoblots are shown ($n = 2$). **c**, Immunofluorescence staining showing PDGF-dependent co-localization of PDGFR and Prx II. HepG2 cells stably expressing PDGFR-β wild type were treated with or without PDGF for 15 min and immunostained with anti-PDGFR and affinity-purified Prx II antibodies. **d**, Subcellular fractionation of NIH3T3 cells disrupted in hypotonic solution. Soluble supernatant (S) and membranous pellet (M) from ultracentrifugation were immunoblotted. **e**, **f**, Protein tyrosine phosphatase (PTPase) activities in soluble supernatants (S) and/or membranous pellets (M) prepared from NIH3T3 cells (**e**) and Prx $II^{-/-}$ MEFs (**f**) infected with control (C) or Prx II wild type (II-WT) retrovirus. Data are represented as means ± s.d. of values from three independent experiments (asterisk, $P < 0.02$ in **e** and $P < 0.001$ in **f**).

© 2005 Nature Publishing Group

smooth muscle cells during vascular remodelling[1,16]. Therefore, we examined *in vitro* whether selective regulation of the PDGFR-β–PLC-γ1 pathway by Prx II also occurs in vascular smooth muscle cells (VSMCs). Modest expression of wild-type Prx II (approximately three times higher than the endogenous level) in human aortic VSMCs clearly and selectively suppressed phosphorylation of both PLC-γ1 and PDGFR at Y579/581 and Y857 (data not shown and Fig. 4a), as well as suppressing chemotactic migration and adhesion induced by PDGF (Fig. 4b). Similarly, when VSMCs cultured from the aorta of wild type and Prx II$^{-/-}$ mice were treated with PDGF, site-selective hyper-phosphorylation of PDGFR-β was observed in Prx II$^{-/-}$ VSMCs in comparison to wild-type cells (Fig. 4c). We then performed an *in vivo* experiment examining the neointimal accumulation of smooth muscle cells during restenosis, which directly represents their proliferation and migration triggered by PDGF[17]. The carotid arteries of uninjured wild type and null mice showed no evidence of neointimal thickening (data not shown). When wire-injured carotid arteries of wild type and Prx II$^{-/-}$ mice were compared, neointimal layers of Prx II$^{-/-}$ mice were significantly thicker than those of wild-type mice, albeit regionally so (Fig. 4d). This neointimal thickening was then abolished by administration of anti-PDGF antibody (Fig. 4e; see also Supplementary Fig. 7). Moreover, immunofluorescence staining showed that Y857, but not Y740, phosphorylation was significantly increased in the injured carotid arteries of Prx II$^{-/-}$ mice (Fig. 4f, g), suggesting that site-selective hyper-phosphorylation of PDGFR occurs *in vivo*. Collectively, these results suggest that loss of Prx II function is sufficient to influence the growth and migration of smooth muscle cells during vascular remodelling.

Our studies have revealed two new signalling mechanisms: (1) endogenous H_2O_2 mediates the site-selective amplification of PDGFR phosphorylation; and (2) Prx II is a physiological regulator of endogenous H_2O_2 signal. 2-Cys peroxiredoxins do not have any known signature motifs for protein–protein interaction. Nonetheless, we were able to predict that a proline-rich motif (PXXP)-like structure in Prx II might be involved in the interaction with PDGFR,

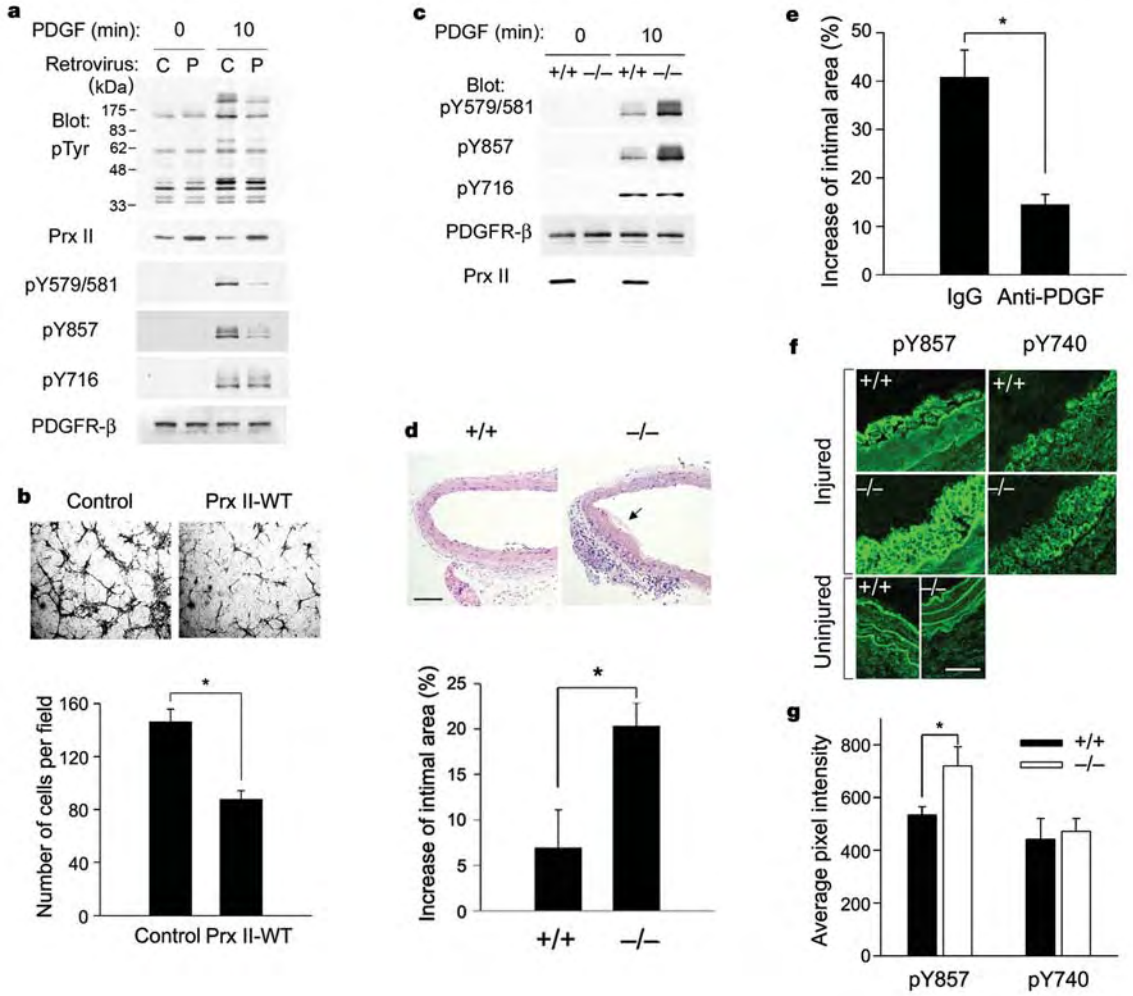


Figure 4 | *In vivo* function of Prx II during restenosis in injured carotid artery. a, Tyrosine phosphorylation of PDGFR-β in human aortic smooth muscle cells. Cells were infected with either control (C) or Prx II wild-type retrovirus (P). **b**, Migratory suppression of human aortic smooth muscle cells by Prx II. A representative picture is shown. Data are means ± s.d. of the number of migrated cells from four random fields, and are representative of two independent experiments (asterisk, $P < 0.001$). **c**, Tyrosine phosphorylation of PDGFR-β in aortic smooth muscle cells prepared from wild type and Prx II$^{-/-}$ mice. **d**, Neointimal thickening in injured carotid arteries of wild type and Prx II$^{-/-}$ mice. A representative picture is shown. The arrow indicates neointimal accumulation of smooth muscle cells. Data (means ± s.e.m.; $n = 5$; asterisk, $P < 0.01$) in the lower graph represent the percentage of neointimal plaque area versus medial layer area. **e**, Neointimal thickening in injured carotid arteries of Prx II$^{-/-}$ mice injected with control IgG or anti-PDGF antibody. Data are presented ($n = 4$; asterisk, $P < 0.001$) as in **d**. **f**, Immunofluorescence images of phospho-PDGFR-β in injured carotid arteries of wild type and Prx II$^{-/-}$ mice. pY857 staining in uninjured carotid arteries is shown as a negative control. Representative images are shown. **g**, Quantification of green fluorescence in neointimal area. Data are means ± s.e.m. of average pixel intensity ($n = 3$; asterisk, $P < 0.005$). Scale bar: 100 μm (**d**); 20 μm (**f**).

© 2005 Nature Publishing Group

as suggested by the interaction of Prx I with the SH3 domain of c-Abl kinase. Recruitment of Prx II to the receptor is thought to allow oxidatively inactivated PTPases, particularly the membrane-associated PTPases, to be reactivated by removing endogenous H_2O_2. Although site specificity of PTPase remains to be demonstrated, other research groups are currently re-investigating the specificity of some PTPases with regard to PDGFR dephosphorylation[18].

It should be noted that our studies demonstrate upregulation of PLC-γ1 as an intermediate signalling component. However, we still cannot exclude the possibility that hyper-activation of signalling molecules other than PLC-γ1 downstream of PDGFR is also involved. Indeed, in addition to PDGFR and PLC-γ1, several other proteins were hyper-phosphorylated in Prx $II^{-/-}$ MEFs (see Fig. 1b). Immunodepletion experiments confirmed that none of them was either FAK or Shc (data not shown), suggesting that unknown phosphorylated proteins might be involved.

Although Prx II is expressed in most cell types and tissues[7], Prx $II^{-/-}$ mice are healthy and grossly normal except for the splenomegaly phenotype in adults[19]. This indicates that Prx II is dispensable for normal development and growth. Therefore, this study provides the first *in vivo* evidence that Prx II has a functional role in the atherosclerotic or injured lesions of blood vessels.

METHODS

Antibodies. The phospho-peptide antigens (SynPep) used for production of the phospho-specific PDGFR-β antibodies were as follows: pY579, DGHEpYIYVDPMQ; pY716, SAELpYSNALPVG; pY740, SDGGpYMDMSKDE; pY751, ESVDpYVPMLDMK; pY771, ESSNpYMAPYDNY; pY857, RDSNpYISKGSTF; pY1009, SSVLpYTAVQPNE; pY1021, GDNDpYIIPLPDP (where amino acid numbering is based on human PDGFR-β sequence and each phosphorylated tyrosine residue has a prefix 'p'). The phospho-peptides were conjugated with keyhole limpet haemocyanin using glutaraldehyde and injected subcutaneously into rabbits. The rabbit antisera were affinity-purified via sequential chromatography on non-phosphorylated and phosphorylated peptide-conjugated Affigel-15 (Bio-Rad) agarose resins. Phospho-Src antibodies were purchased from Biosource. Phospho-Akt and phospho-ERK antibodies were from Cell Signalling Technology. c-Src, Akt, PDGFR-β (M−20) and ERK2 were from Santa Cruz Biotechnology. PLC-γ1 and phospho-PLC-γ1 (pY783 and pY1253) antibodies have been described[11].

Primary culture and retroviral infection of MEFs and VSMCs. Three independent clonal lines of wild type and Prx $II^{-/-}$ MEFs were prepared and cultured as described elsewhere[20]. All experiments were performed with the MEFs at the fourth passage. Human aortic smooth muscle cells were cultured according to the manufacturer's protocol (Clonetics). The aortic smooth muscle cells from 10-week-old wild type and Prx $II^{-/-}$ mice were prepared as described[21]. The cells were subcultured up to four passages and examined by immunostaining with an antibody to α-smooth muscle actin (Sigma). The primary cells were serum starved for 24 h with 0.5% FBS before PDGF-BB (25 ng ml^{-1}, Upstate) stimulation.

Control and Prx II retrovirus were prepared from a transfected PT67 packaging cell line[20] or from a transiently transfected Phoenix-Ampho packaging cell line (http://www.stanford.edu/group/nolan/index.html). MEFs, VSMCs and NIH3T3 were infected with the retrovirus at a multiplicity of infection (M.O.I.) of 10 for 24 h before serum starvation.

Protein tyrosine phosphatase assay. The control or Prx II wild-type retrovirus-infected NIH3T3 (5×10^5 cells) or Prx $II^{-/-}$ MEFs (1×10^6 cells) were serum starved for 24 h and stimulated with or without PDGF-BB for 10 min. After stimulation, the cells were frozen in liquid nitrogen and moved to an anaerobic chamber. The cells were scraped in hypotonic buffer (2 mM $MgCl_2$, 25 mM KCl, 0.5 mM EDTA, 0.5 mM EGTA and a protease inhibitor cocktail in a 50 mM PIPES buffer, pH 6.5) containing 10 mM iodoacteate and 10 mM *N*-ethylmaleimide. The homogenates were obtained by passing through a 26-guage needle 15 times and then incubated for a further 20 min in order to achieve complete alkylation of free thiols. The labelling was quenched by adding 50 mM DTT. Then, the homogenates were then centrifuged at 100,000*g* to separate the soluble supernatant and insoluble membranous pellet. The membranous pellets were dissolved in lysis buffer (1% Triton X-100, 150 mM NaCl, 10% glycerol, 1 mM EDTA, 2 mM EGTA, in 20 mM HEPES buffer, pH 7.0) with ultrasonication. The proteins were measured using a bicinchoninic acid assay. The samples were diluted in a 10 mM DTT-containing reaction buffer and the PTPase activities were measured in a 96-well plate coated with poly-(Glu_4-pTyr) peptides, according to the manufacturer's protocol (Universal tyrosine phosphatase assay kit, Takara Bio., MK-411). PTPase activity was quantified from a standard curve obtained using the recombinant CD45 tyrosine phosphatase and is expressed as milliunits per milligram protein used for the assay.

Carotid artery injury and immunofluorescence staining. A transluminal wire injury to the left common carotid artery was performed as described[22]. Ten-week-old male mice were anaesthetized, the left external carotid artery was exposed, and its branches were electro-coagulated. A 0.016-inch flexible angioplasty guide wire was pushed 1 cm in through the transverse arteriotomy of the external carotid artery, and the endothelial denudation was achieved by six passes along the common carotid artery. One week after the injury, the common carotid arteries were excised after transcardiac perfusion-fixation with heparinized saline containing 3.7% formaldehyde, and paraffin embedded. Five serial tissue sections (100-μm interval and 3-μm thickness) were obtained from the middle area of common carotid arteries. Each slide was stained with haematoxylin and eosin. Cross-sectional areas of medial and neointimal layers were measured and analysed using a computerized imaging system (MetaMorpho imaging v5.0 (Molecular Devices) and CoolSNAP camera system with BX51 microscopy (Olympus)).

Tissue sections were also used for immunostaining phospho-PDGFR-β (each antibody at 1:50 dilution). Immunofluorescence images were taken using a LSM510 confocal laser-scanning microscope (Zeiss). Three areas in the neointimal layer (100 μm^2) were selected and the average fluorescence pixel intensity was determined.

For neutralizing the plasma PDGF, control IgG or anti-PDGF antibody (catalogue number AF-220-NA, R&D systems) was injected into Prx $II^{-/-}$ mice three times during the course of 1 week before and after the injury (intravenously at −1 and 0 day, subcutaneously at +4 day; 5 μg IgG per approximately 20 g of body weight at a time).

Statistics. Data were analysed by a Student's *t*-test on SigmaPlot 8.0 software and statistical significance (*P*-value) was determined.

Received 18 January; accepted 29 March 2005.

1. Heldin, C. H. & Westermark, B. Mechanism of action and *in vivo* role of platelet-derived growth factor. *Physiol. Rev.* **79**, 1283–1316 (1999).
2. Sundaresan, M., Yu, Z. X., Ferrans, V. J., Irani, K. & Finkel, T. Requirement for generation of H_2O_2 for platelet-derived growth factor signal transduction. *Science* **270**, 296–299 (1995).
3. Bae, Y. S. *et al.* Platelet-derived growth factor-induced H_2O_2 production requires the activation of phosphatidylinositol 3-kinase. *J. Biol. Chem.* **275**, 10527–10531 (2000).
4. Kang, S. W. *et al.* Mammalian peroxiredoxin isoforms can reduce hydrogen peroxide generated in response to growth factors and tumor necrosis factor-alpha. *J. Biol. Chem.* **273**, 6297–6302 (1998).
5. Rhee, S. G., Kang, S. W., Chang, T. S., Jeong, W. & Kim, K. Peroxiredoxin, a novel family of peroxidases. *IUBMB Life* **52**, 35–41 (2001).
6. Wood, Z. A., Schroder, E., Robin Harris, J. & Poole, L. B. Structure, mechanism and regulation of peroxiredoxins. *Trends Biochem. Sci.* **28**, 32–40 (2003).
7. Chae, H. Z., Kim, H. J., Kang, S. W. & Rhee, S. G. Characterization of three isoforms of mammalian peroxiredoxin that reduce peroxides in the presence of thioredoxin. *Diabetes Res. Clin. Pract.* **45**, 101–112 (1999).
8. Kim, H. K. *et al.* PDGF stimulation of inositol phospholipid hydrolysis requires PLC-γ1 phosphorylation on tyrosine residues 783 and 1254. *Cell* **65**, 435–441 (1991).
9. Baxter, R. M., Secrist, J. P., Vaillancourt, R. R. & Kazlauskas, A. Full activation of the platelet-derived growth factor β-receptor kinase involves multiple events. *J. Biol. Chem.* **273**, 17050–17055 (1998).
10. Rhee, S. G. & Bae, Y. S. Regulation of phosphoinositide-specific phospholipase C isozymes. *J. Biol. Chem.* **272**, 15045–15048 (1997).
11. Sekiya, F., Poulin, B., Kim, Y. J. & Rhee, S. G. Mechanism of tyrosine phosphorylation and activation of phospholipase C-γ1. Tyrosine 783 phosphorylation is not sufficient for lipase activation. *J. Biol. Chem.* **279**, 32181–32190 (2004).
12. Saito, S. *et al.* Ligand-independent trans-activation of the platelet-derived growth factor receptor by reactive oxygen species requires protein kinase C-δ and c-Src. *J. Biol. Chem.* **277**, 44695–44700 (2002).
13. Rosado, J. A. *et al.* Hydrogen peroxide generation induces pp60src activation in human platelets: evidence for the involvement of this pathway in store-mediated calcium entry. *J. Biol. Chem.* **279**, 1665–1675 (2004).
14. Yang, K. S. *et al.* Inactivation of human peroxiredoxin I during catalysis as the result of the oxidation of the catalytic site cysteine to cysteine-sulfinic acid. *J. Biol. Chem.* **277**, 38029–38036 (2002).
15. Rhee, S. G., Bae, Y. S., Lee, S. R. & Kwon, J. Hydrogen peroxide: a key messenger that modulates protein phosphorylation through cysteine oxidation. *Sci STKE* **2000**, PE1 (2000).
16. Ross, R. The pathogenesis of atherosclerosis: a perspective for the 1990s. *Nature* **362**, 801–809 (1993).

© 2005 Nature Publishing Group

17. Ferns, G. A. *et al.* Inhibition of neointimal smooth muscle accumulation after angioplasty by an antibody to PDGF. *Science* **253,** 1129–1132 (1991).
18. Persson, C. *et al.* Site-selective regulation of platelet-derived growth factor beta receptor tyrosine phosphorylation by T-cell protein tyrosine phosphatase. *Mol. Cell. Biol.* **24,** 2190–2201 (2004).
19. Lee, T. H. *et al.* Peroxiredoxin II is essential for sustaining life span of erythrocytes in mice. *Blood* **101,** 5033–5038 (2003).
20. Kang, S. W. *et al.* Cytosolic peroxiredoxin attenuates the activation of Jnk and p38 but potentiates that of Erk in Hela cells stimulated with tumor necrosis factor-α. *J. Biol. Chem.* **279,** 2535–2543 (2004).
21. Ohmi, K. *et al.* A novel aortic smooth muscle cell line obtained from p53 knock out mice expresses several differentiation characteristics. *Biochem. Biophys. Res. Commun.* **238,** 154–158 (1997).
22. Schober, A. *et al.* Stabilization of atherosclerotic plaques by blockade of macrophage migration inhibitory factor after vascular injury in apolipoprotein E-deficient mice. *Circulation* **109,** 380–385 (2004).
23. Woo, H. A. *et al.* Reversible oxidation of the active site cysteine of peroxiredoxins to cysteine sulfinic acid. Immunoblot detection with antibodies specific for the hyperoxidized cysteine-containing sequence. *J. Biol. Chem.* **278,** 47361–47364 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. B. Kwon, E. S. Oh and T. H. Lee for reagents and discussions; J. Kim and I. C. Baines for critically reading the manuscript; G. P. Nolan for pBMN retroviral plasmids; A. Kazlauskas for providing HepG2 cells overexpressing PDGFR-β wild type; and D. J. Lee for immunofluorescence staining. We also thank I. H. Lee, H. I. Park and researchers in Labfrontier Life Science Institute for technical assistance. This work was supported by Korea Science and Engineering Foundation through the Center for Cell Signalling Research in Ewha Womans University and by a grant from the Functional Proteomics Center of the 21st Century Frontier Research Program funded by the Ministry of Science and Technology of the Korean government. M.H.C. is supported by a Brain Korea 21 grant from the Ministry of Education and Human Resources Development.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.W.K. (kangsw@ewha.ac.kr) or S.G.R. (sgrhee@nih.gov).

© 2005 Nature Publishing Group

LETTERS

The Mesp2 transcription factor establishes segmental borders by suppressing Notch activity

Mitsuru Morimoto[1], Yu Takahashi[3], Maho Endo[1] & Yumiko Saga[1,2]

The serially segmented (metameric) structures of vertebrates are based on somites that are periodically formed during embryogenesis. A 'clock and wavefront' model has been proposed to explain the underlying mechanism of somite formation[1], in which the periodicity is generated by oscillation of Notch components (the clock) in the posterior pre-somitic mesoderm (PSM)[2–6]. This temporal periodicity is then translated into the segmental units in the 'wavefront'[7,8]. The wavefront is thought to exist in the anterior PSM and progress backwards at a constant rate; however, there has been no direct evidence as to whether the levels of Notch activity really oscillate and how such oscillation is translated into a segmental pattern in the anterior PSM. Here, we have visualized endogenous levels of Notch1 activity in mice, showing that it oscillates in the posterior PSM but is arrested in the anterior PSM. Somite boundaries formed at the interface between Notch1-activated and -repressed domains. Genetic and biochemical studies indicate that this interface is generated by suppression of Notch activity by mesoderm posterior 2 (Mesp2) through induction of the lunatic fringe gene (*Lfng*). We propose that the oscillation of Notch activity is arrested and translated in the wavefront by Mesp2.

Mesp2—a basic helix–loop–helix-type transcription factor—and its related proteins have an essential role in both segmentation and rostro-caudal patterning of somites in the anterior PSM[9–12]. To understand further the involvement of Mesp2 in boundary formation, we generated a *Mesp2–venus* knock-in mouse (Supplementary Fig. 1) in which Mesp2 localization is detected in live embryos using confocal microscopy (Fig. 1a–g). As expected, the Mesp2–venus fusion protein was detected in the nuclei of cells (Fig. 1d). Furthermore, in embryos at 8.75 days post coitus (d.p.c.), two to three segmental stripes are detectable, showing an anterior to posterior gradient for each segment (Fig. 1a–d). The signal in the most anterior segment is very weak but a second stripe has a clear anterior boundary that perfectly matches the segmental border; indeed, it lines the entire border. The strongest third stripe also has a clear anterior border in the PSM in which no sign of a morphological boundary can be detected. This band may therefore demarcate the next segmental border. In later-stage embryos, one clear stripe was detected, although we occasionally observed an additional signal in the newly formed somite (Fig. 1e– g).

We have previously shown that the expression domain of *Mesp2* transcript changes during somitogenesis[13], appearing at S−1 or S−2 as a single band of one somite in length, and then gradually shrinking, leaving the rostral half of its expression domain intact. To elucidate further the dynamic behaviour of Mesp2, we have recorded its expression pattern using time-lapse microscopy. A new Mesp2 expression domain is evident as a single band of about one somite in length, whereas the expression domain of the pre-existing band undergoes changes such that the rostral part of the band is maintained longer before finally disappearing (Supplementary Movie 1 and Supplementary Fig. 2). This dynamic change is consistent with the expression profile of *Mesp2* messenger RNA, previously obtained using explant cultures[13]. However, because Mesp2–venus is a fusion protein, it may not faithfully reproduce the pattern of endogenous protein distribution. This prompted us to generate antibodies against the Mesp2 protein. The antibody-staining pattern reveals that the distribution of endogenous Mesp2 is similar to that of Mesp2–venus (Fig. 1h–j). The neighbouring section was subjected to *in situ* hybridization analysis. The mRNA localization pattern was found to be consistent with the immunohistochemical results, indicating that message is translated immediately and that the protein is rapidly degraded but the transcript disappears earlier, as expected (Fig. 1j). We occasionally observed two bands of Mesp2 protein: an anterior band located in the segmented border and an posterior one in the next presumptive border. In this case, transcript was only observed in the posterior region (Fig. 1h). Taken together, we conclude that the anterior border of the expression domain of the Mesp2 protein coincides with the next segmental border in the PSM.

The expression of Notch components, such as *Hes7* (hairy and enhancer of split 7) and *Lfng*, oscillates in the PSM[3,4,6], whereas *Notch1* is expressed throughout the entire PSM and at higher levels in the anterior PSM[9]. We attempted to visualize the activation of Notch signalling by detecting a processed NICD (Notch intracellular domain) using a specific antibody. Unlike its RNA expression pattern, Notch1 activity exhibits a dynamic pattern in the PSM during somitogenesis. We analysed 52 embryos, which exhibited four distinct patterns; the positive signals were detected in the tailbud (Fig. 1k), the middle of the PSM (Fig. 1l, m) or more anteriorly within the PSM (Fig. 1n). Notably, in most of the embryos, Notch activation is in a similar phase to the transcript of *Lfng*, which encodes a glycosyltransferase that can modify Notch activity[14,15] (Fig. 1k–n). The expression domain of *Lfng* appears slightly later than Notch activation, supporting the idea that *Lfng* transcription is activated by Notch signalling, as previously indicated[16]. These results indicate that the activity of Notch oscillates in the posterior PSM. In a previous report, Hes7 was shown to function as a suppressor of *Hes7* and *Lfng* transcription[17], forming a feedback loop in the core oscillator. We thus tested the pattern of Notch activation in mice defective in Notch pathway components (Fig. 2a–e). In *Dll1*-null embryos[18], no Notch1 activity is found anywhere in the PSM (Fig. 2c), whereas in *Lfng*-null embryos[19], Notch1 activity is detected throughout the entire PSM but does not seem to oscillate (Fig. 2d). These observations indicate that the Dll1 ligand is required for Notch1 activation, and that Lfng functions as a suppressor of Notch activity, thus generating the oscillatory activation of Notch1.

In the anterior PSM, Notch oscillatory activity is arrested and localizes in the caudal portion of the S0 somite, whereas *Lfng* transcripts are detectable in the rostral part of S − 1 and their

[1]Division of Mammalian Development, National Institute of Genetics, and [2]Sokendai, Yata 1111, Mishima 411-8540, Japan. [3]Cellular & Molecular Toxicology Division, National Institute of Health Sciences, 1-18-1 Kamiyoga, Setagaya-ku, Tokyo 158-8501, Japan.

© 2005 Nature Publishing Group

expression seldom overlaps with the Notch1 activation domain (Fig. 1k–n). Of note, Notch activation and *Lfng* expression domains almost co-occur in the posterior PSM, but are segregated into the rostral and caudal parts in the anterior PSM. This suggests that *Lfng* expression is activated by an additional factor in the anterior PSM. Mesp2 may well be this factor, as it is co-expressed with *Lfng* in this region (refs 20, 21; see also Supplementary Fig. 3). We propose that in the anterior PSM, Mesp2 may downregulate Notch activity through the activation of Lfng, which is a Notch modulator. To address this possibility, we first examined the relationship between Mesp2 and Notch activity in the anterior PSM. Double staining with anti-active-NICD and anti-Mesp2 antibodies enabled us to examine the spatial relationship between the localization of these two factors in a single section (Fig. 2f–u). At the initial phase of Mesp2 expression the two expression domains partially overlap in such a way that cells in the posterior Notch activation domain co-express Mesp2 (Fig. 2f–i). The overlapping domain gradually reduces as expression levels of Mesp2 increase (Fig. 2j–m). Ultimately, the two expression domains are completely separated from each other and form a clear boundary (Fig. 2n–q). Notably, the strongest detectable signals of both Notch activity and Mesp2 are found in close proximity to the boundary on both sides. This boundary will form the next segmental border, as indicated by the expression of Mesp2–venus (Fig. 1a–g). Once a new border is generated, next Notch activity appears in the anterior PSM and Mesp2 protein appears just in the middle of the Notch activation domain, to generate a next segmental border (Fig. 2r–u).

We next examined whether downregulation of Notch activity by Mesp2 is mediated by Lfng. In *Lfng*-null embryos[19], Mesp2 protein is

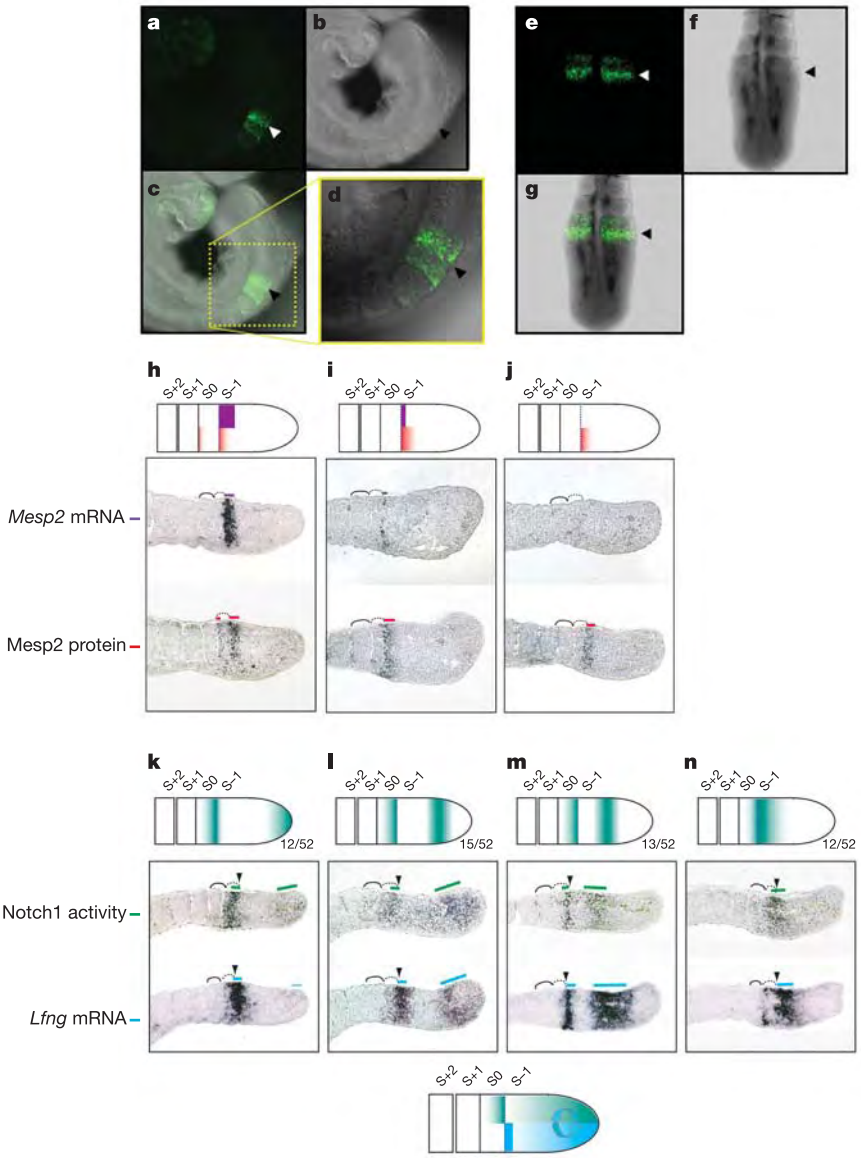


Figure 1 | Visualization of the Mesp2 protein and oscillation of Notch1 activity and *Lfng* transcripts. **a–g**, Expression of the Mesp2–venus fusion protein is shown in 8.75 d.p.c. (**a–d**) and 10.5 d.p.c. (**e–g**) embryos as confocal images (**a**, **e**; green), bright-field images (**b**, **f**) and merged images (**c**, **g**). The boxed area in **c** is shown at higher magnification in **d**. Stronger Mesp2–venus signals are localized at the anterior border of the predicted next somite (next borders are indicated by arrowheads). **h–j**, *Mesp2* transcript and protein (using anti-Mesp2 antibody) expression was compared using three sets of serial sections of 11.5 d.p.c. embryos. **k–n**, Notch1 activity periodically oscillating in the PSM may activate *Lfng* transcription. Notch1 activity was revealed by immunostaining with antibodies against active-NICD and the adjacent sections were stained by *in situ* hybridization with *Lfng* probe. Solid and dotted lines indicate S+1 and S0 somite regions, respectively, and coloured bars represent the stained domain. $n = 52$ embryos. Each expression pattern is summarized on the top of each panel.

© 2005 Nature Publishing Group

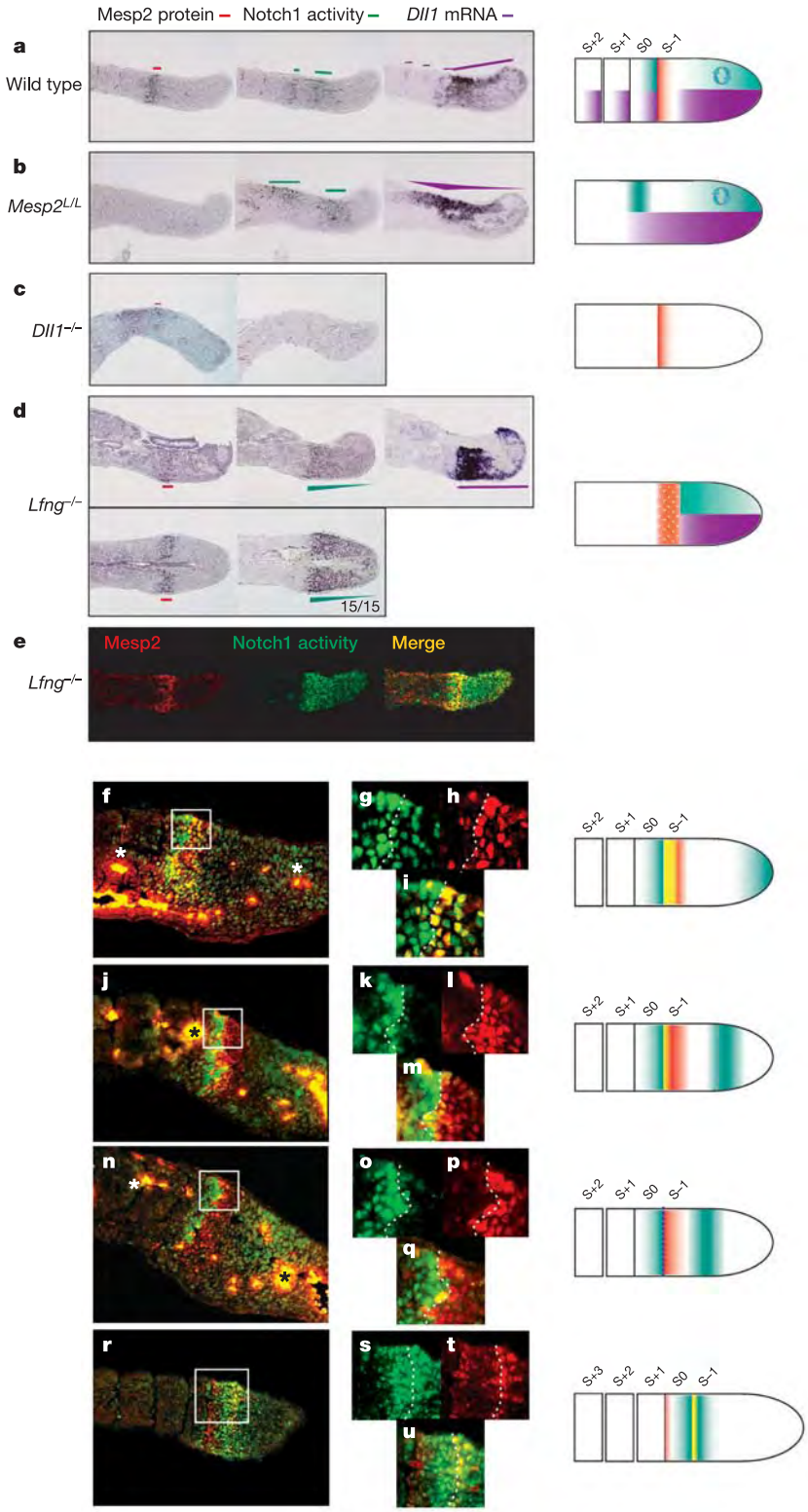


Figure 2 | Analyses of gene and protein expression critical for clock oscillation and its arrest. **a**–**e**, Serial sections were prepared from 11.5 d.p.c. wild type (**a**), *Mesp2*$^{L/L}$ (**b**), *Dll1*$^{-/-}$ (**c**) or *Lfng*$^{-/-}$ (**d**, **e**) embryonic tails. Each section was stained with anti-Mesp2 or anti-active-NICD antibodies, or with *Dll1* or *Lfng* RNA probe. Coloured bars represent stained domains. In *Mesp2*$^{L/L}$ embryos, Notch1 activity is expanded in the anterior PSM, and expression of *Dll1* is also strongly expanded. In *Dll1*$^{-/-}$ embryos, expression of Mesp2 is decreased and Notch1 activity is almost completely lost. In all *Lfng*$^{-/-}$ embryos ($n = 15$), Notch1 activity is strongly detected throughout the PSM but its oscillatory expression is lost. Each expression pattern is summarized in the right panels. **f**–**u**, Mesp2 protein (red) and Notch1 activity (green) were compared by double immunostaining. A merged set of four images representing different patterns is shown (**f**, **j**, **n**, **r**). The border regions (boxed areas) are magnified and the signals are shown either separately (**g**, **h**, **k**, **l**, **o**, **p**, **s**, **t**) or merged (**i**, **m**, **q**, **u**). Dotted lines indicate the boundaries, and the expression patterns are summarized in the right panels. Asterisk indicates nonspecific immunofluorescence of blood cells.

© 2005 Nature Publishing Group

expressed in the appropriate location, although the expression is broad and does not show a normal anterior–posterior gradient (Fig. 2d). Hence, Lfng is required for normal Mesp2 protein localization. Importantly, the anterior limit of the Notch1 activation domain, which encompasses the PSM and does not have oscillatory Notch1 activity, overlaps with the Mesp2 expression domain, indicating that Mesp2 fails to suppress Notch activity and thus the Notch activation domain does not segregate in *Lfng*-null embryos (Fig. 2d, e).

To study further the function of Mesp2 in the control of both Notch activity and *Lfng* expression, a number of neighbouring sections of *Mesp2*-null embryos were subjected to NICD and *Lfng* staining. We observed that Notch1 activity and *Lfng* expression oscillate normally in the posterior PSM in *Mesp2*-null embryos; however, their expression is abnormal and expands anteriorly in the anterior PSM, and they do not make narrow stripes (Fig. 3a). This was more clearly shown in whole-mount staining of *Lfng* expression in 9.5 d.p.c. embryos (Supplementary Fig. 4). The expanded signals could be sustained by ectopic expression of *Dll1*, which extends anteriorly in *Mesp2*-null embryos (Fig. 2b and ref. 9). More importantly, the two domains of Notch1 activation and *Lfng* expression are never segregated and stay in unison (compare Figs 1k–m and 3a). We interpret this to mean that in wild-type embryos, Mesp2 disrupts the relationship between Notch activity and Lfng by stably activating Lfng, a Notch repressor, in the anterior PSM. In order to determine whether Mesp2 acts on the *Lfng* enhancer and regulates its gene expression, we generated a reporter construct that contained a 2.3-kilobase (kb) upstream enhancer sequence (defined by transgenic analyses reported in refs 20 and 21) linked to a luciferase gene (Fig. 3b). As shown in Fig. 3c, both Mesp2 and NICD act as activators in this reporter system, and a synergistic effect was also observed. However, these activities are strongly suppressed by Hes7, which is active in the posterior PSM and is downregulated in the anterior PSM[22]. Furthermore, we found that Mesp2 specifically binds to a portion of the *Lfng* enhancer that contains a putative N-box (Fig. 3d). These results strongly suggest that *Lfng* is activated by NICD and suppressed by Hes7 in the posterior PSM, whereas it is activated by Mesp2 and hence its expression is restricted and confined to the Mesp2 expression domain, thereby inhibiting and arresting the oscillation of Notch activity (Fig. 4 and ref. 16). This

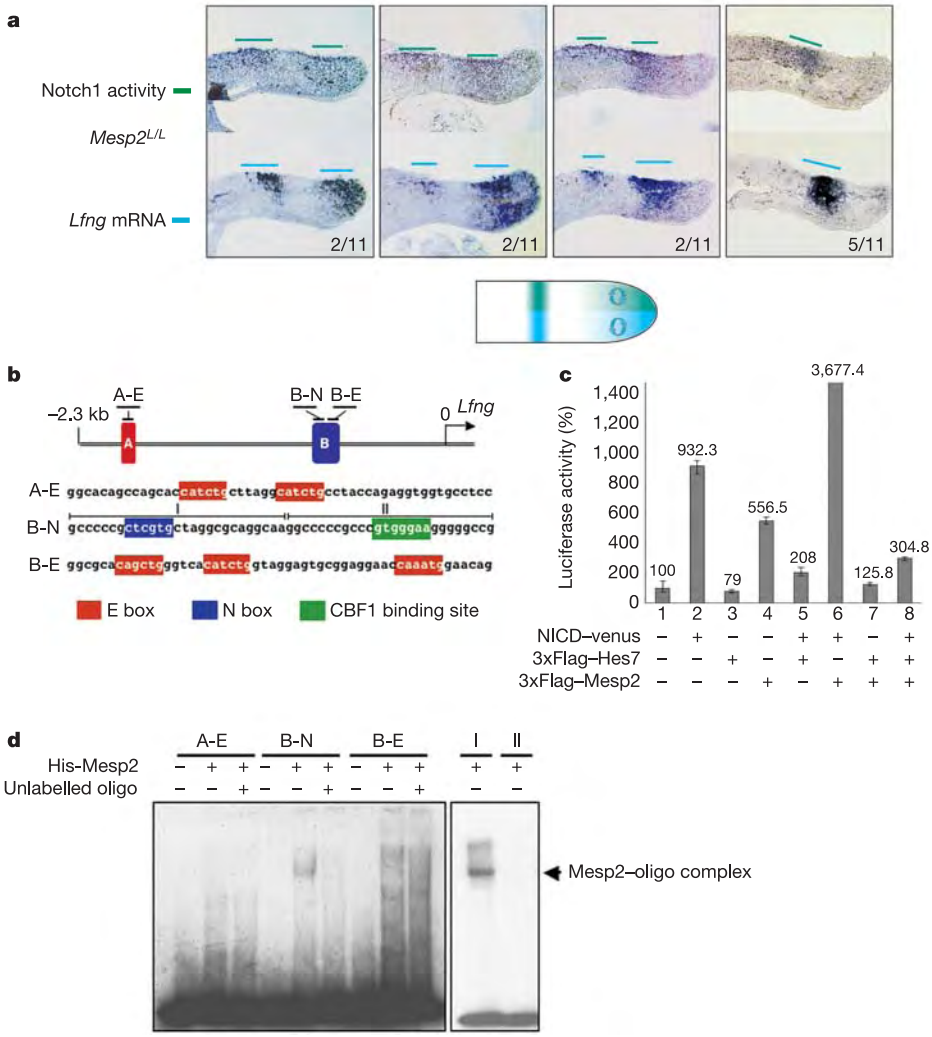


Figure 3 | Mesp2 may activate *Lfng* to arrest the oscillation of Notch activity in the anterior PSM. a, Serial sections of mouse tail regions were prepared from *Mesp2*$^{L/L}$ embryos ($n = 11$) and stained with either anti-active-NICD antibody (top) or *Lfng* probe (bottom). Notch1 activity and *Lfng* signal oscillate in the posterior PSM but do not arrest in the anterior PSM. **b**, A schematic representation of a luciferase reporter construct with an *Lfng* 5′ upstream region (top). DNA fragments A-E, B-N and B-E are used for gel mobility shift assays. **c**, Luciferase reporter assay showing that NICD and Mesp2 are strong activators of the *Lfng* enhancer, whereas Hes7 is a strong repressor of the *Lfng* enhancer. **d**, Mesp2 specifically binds only to the N-box-containing B-N fragment.

© 2005 Nature Publishing Group

result is consistent with previous studies indicating that the regulatory region of *Lfng* is separable into a region A that regulates the cyclic expression in the posterior PSM, and a region B that regulates expression in the anterior PSM[20,21].

The PSM can be divided into two regions based on the finding that the clock oscillates in the posterior PSM whereas it is slowed down and arrested in the anterior PSM. A switch between these two states is thought to be triggered by FGF signalling[23,24], and Mesp2 is crucial in this process. In the absence of Mesp2, the Dll1-dependent relationship between Notch activity and Lfng is maintained; that is, the normal wavefront is not generated, even if the clock system functions normally. We have previously shown that a major role of Mesp2 is to generate rostro-caudal polarity within the somite through suppression of the Dll1-dependent Notch pathway in the presumptive rostral domain[13]. We now propose that Mesp2 has an additional role in segment border formation by restricting *Lfng* expression in the anterior PSM. In both cases, Mesp2 functions as a strong suppressor of Notch activity.

We have shown that Lfng is a negative regulator of Dll1–Notch signalling. The results are consistent[16] or contrast[25,26] with previous findings. In the *Drosophila* wing disc, Fringe inhibits Serrate-dependent but potentiates Delta-dependent Notch signalling, whereas in somite segmentation, Lfng suppresses Dll1-dependent Notch activation (Fig. 4 and ref. 16). In a mammalian cell culture system, Dll1–Notch signalling is enhanced by Lfng[27]. These discrepancies could be due to different cellular contexts and different biological systems, emphasizing variation of modulating mechanisms for Notch signalling.

We have now provided direct evidence that Notch activity constitutes a core of the segmentation clock and that Mesp2 arrests the oscillation of Notch activity and initiates a segmentation programme. The next crucial question would be the mechanism by which Mesp2 is turned on in the anterior PSM. Our analyses have revealed a strict spatial relationship between stabilized Notch activation domains and the emergence of Mesp2 expression (Fig. 2f–u; see also Supplementary Fig. 5). In the absence of Notch signalling, however, the Mesp2 expression region is approximately normal in the anterior PSM (Fig. 2c), and thus its initial activation depends on positional information such as the Fgf8 (refs 23, 24) and retinoic acid (RA) gradients[28]. However, the role of RA remains elusive because no severe effect on Mesp2 expression is observed in a targeted disruption of CYP26 (ref. 29), a degrading enzyme of RA (Supplementary Fig. 6). Detailed enhancer analysis of Mesp2 will provide further insights into the mechanism of where and how a sharp NICD–Mesp2 boundary forms.

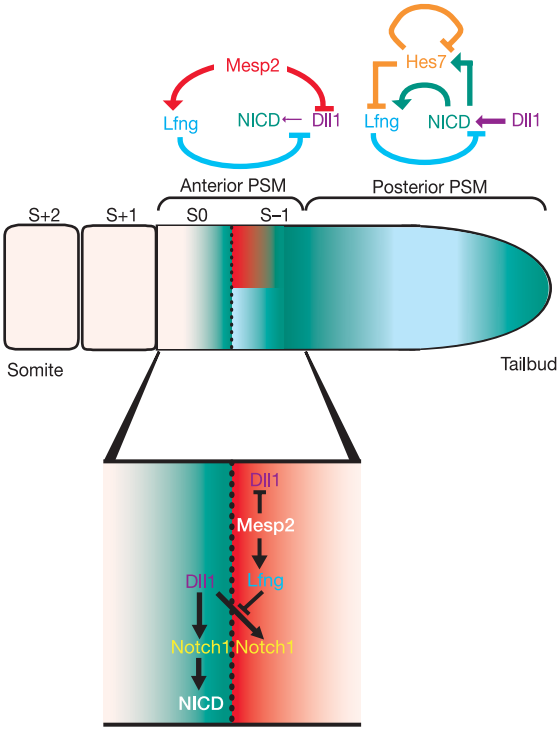


Figure 4 | Schematic representation of the regulatory mechanism underlying the clock system and of the implications of Mesp2 function in establishing the segmental boundary. In the posterior PSM, the Dll1–Notch signal initially activates both *Lfng* and *Hes7*. Hes7 is a strong transcriptional repressor of *Lfng* and of the *Hes7* gene itself, whereas Lfng is a negative modulator of the Notch receptor in this cellular context. The positive and negative feedback loops thus generates oscillation of Notch1 activity. In the anterior PSM, the Notch1 activity and *Lfng* waves are no longer subject to negative regulation by Hes7, as Mesp2 now becomes a major regulator of *Lfng* activation and *Dll1* suppression. As a result, both Notch1 activity and *Lfng* waves are arrested and generate a clear boundary between the Notch1 activity domain and the Mesp2 expression domain, which produce the next segmental boundary.

METHODS

Time-lapse recording using confocal microscopy. Heterozygous Mesp2–venus mouse embryos were collected at 10.5 d.p.c. The tail region was excised and cultured with DMEM/F12 medium (without Phenol red) containing 10% fetal calf serum on a 35-mm glass-bottomed dish (Matsunami Glass) at 37 °C. For the culturing of embryos, the Heating Insert P and Incubator S systems (Zeiss) were used to obtain optimal growth conditions. Scanning for the fluorescence protein venus[30] (excitation at 514 nm) and bright imaging of the tail region was performed with an Axiovert 200M confocal laser scanning microscope (Zeiss). Three-dimensional (x, y, t) image data sets were recorded at 10-min intervals as typical time-lapse imaging sessions ($n = 30$). The colour images were captured using an Axiocam (Zeiss) and converted to pseudo-colours using Zeiss Axiovision software.

***In situ* hybridization and immunohistochemistry.** The tails were dissected at 10.5–11.5 d.p.c., fixed with 4% PFA and embedded in OCT. For *in situ* hybridizations, frozen sections (6–8 μm) were hybridized with digoxigenin-labelled antisense cRNA probes (Roche). Hybridized probes were detected using AP-conjugated anti-DIG sheep antibody (1:1,000, Roche) and visualized with NBT/BCIP (1:50, Roche).

For immunohistochemistry analysis, frozen sections (6–8 μm) were immersed in unmasking solution (Vector Laboratories) and autoclaved at 105 °C for 15 min to enable antigen retrieval. Immunostaining was then performed following a standard method with either anti-Mesp2 (1:500) or anti-NICD (1:500, Cell signaling technology) as primary antibodies, biotinylated anti-rabbit IgG goat antibody (1:200, Vector Laboratories) as a secondary antibody and an ABC kit (Vector Laboratories) for signal detection.

For double-immunofluorescent staining, antibody reactions and the detection of Notch1 activity and Mesp2 were separately conducted after antigen retrieval. Detection of Notch1 activity was performed first with anti-active-NICD (1:200, Cell signaling technology) primary antibody, horseradish peroxidase-conjugated anti-rabbit IgG donkey antibody (1:100, Amersham Pharmacia Biotech) secondary antibody and fluorescein isothiocyanate-conjugated Tyramid (Perkin Elmer) signal detection. Detection of Mesp2 was then performed with anti-Mesp2 (1:200) biotinylated anti-rabbit IgG goat antibody (1:200, Vector Laboratories) and Alexa Flour-594-conjugated Streptavidin (Molecular Probes). The corresponding signals were observed with an Olympus BX61 fluorescence microscope system using an ORCA-ER digital camera (HAMAMATSU photo) and analysed with MetaMorph software (Universal Imaging).

Luciferase assay. For luciferase reporter analysis under the control of the *Lfng* upstream 2.5-kb (*Eco*RI–*Not*I) enhancer (0.2 μg), the reporter construct was transfected with or without the expression vectors for NICD–venus (50 ng), 3 × Flag-tagged Hes7 (20 ng) or 3 × Flag-tagged Mesp2 (50 ng) into NH3T3 cells (0.25×10^5 cells per well in 24-multiwell plates) using Lipofectamine Plus (Invitrogen) following the manufacturer's instructions. The vector containing the Renilla luciferase gene under the control of the thymidine kinase promoter (10 ng) was co-transfected as an internal standard to normalize for transfection efficiency. After 36 h, luciferase activities were measured using a Dual Luciferase Assay Kit (Promega).

Gel-shift assay. A 6 × His-Mesp2 protein was produced using a baculovirus

© 2005 Nature Publishing Group

expression system with Sf-9 cells, and then purified through a TALON metal affinity resin (Clontech) with 100 mM imidazole as the elution buffer. The double-stranded DNA oligonucleotide probes were end-labelled with DIG and protein–DNA complexes were detected using a Roche DIG gel shift kit. Binding reactions were carried out for 30 min on ice, and protein–DNA complexes were analysed on 6% native-polyacrylamide gels. The sense strand sequences of the oligonucleotide probes used were: A-E 5′-GGGCACAGCCAGCACCATCTGCT-TAGGCATCTGCCTACCAGAGGTGGTGCCTCC-3′; B-N 5′-GCCCCCGCT CGTGCTAGGCGCAGGCAAGGCCCCCGCCCGTGGGAAGGGGGCCG-3′; B-E 5′-GGCGCACAGCTGGGTCACATCTGGTAGGAGTGCGGAGGAACC AAATGGAACAG-3′.

Received 3 January; accepted 31 March 2005.

1. Cooke, J. & Zeeman, E. C. A clock and wavefront model for control of the number of repeated structures during animal morphogenesis. *J. Theor. Biol.* **58,** 455–476 (1976).
2. Palmeirim, I., Henrique, D., Ish-Horowicz, D. & Pourquie, O. Avian hairy gene expression identifies a molecular clock linked to vertebrate segmentation and somitogenesis. *Cell* **91,** 639–648 (1997).
3. McGrew, M. J., Dale, J. K., Fraboulet, S. & Pourquie, O. The lunatic fringe gene is a target of the molecular clock linked to somite segmentation in avian embryos. *Curr. Biol.* **8,** 979–982 (1998).
4. Forsberg, H., Crozet, F. & Brown, N. A. Waves of mouse Lunatic fringe expression, in four-hour cycles at two-hour intervals, precede somite boundary formation. *Curr. Biol.* **8,** 1027–1030 (1998).
5. Jiang, Y. J. *et al.* Notch signalling and the synchronization of the somite segmentation clock. *Nature* **408,** 475–479 (2000).
6. Bessho, Y. *et al.* Dynamic expression and essential functions of Hes7 in somite segmentation. *Genes Dev.* **15,** 2642–2647 (2001).
7. Saga, Y. & Takeda, H. The making of the somite: molecular events in vertebrate segmentation. *Nature Rev. Genet.* **2,** 835–845 (2001).
8. Pourquie, O. The segmentation clock: converting embryonic time into spatial pattern. *Science* **301,** 328–330 (2003).
9. Saga, Y., Hata, N., Koseki, H. & Taketo, M. M. Mesp2: a novel mouse gene expressed in the presegmented mesoderm and essential for segmentation initiation. *Genes Dev.* **11,** 1827–1839 (1997).
10. Buchberger, A., Seidl, K., Klein, C., Eberhardt, H. & Arnold, H. H. cMeso-1, a novel bHLH transcription factor, is involved in somite formation in chicken embryos. *Dev. Biol.* **199,** 201–215 (1998).
11. Sparrow, D. B. *et al.* Thylacine 1 is expressed segmentally within the paraxial mesoderm of the *Xenopus* embryo and interacts with the Notch pathway. *Development* **125,** 2041–2051 (1998).
12. Sawada, A. *et al.* Zebrafish Mesp family genes, *mesp-a* and *mesp-b* are segmentally expressed in the presomitic mesoderm, and Mesp-b confers the anterior identity to the developing somites. *Development* **127,** 1691–1702 (2000).
13. Takahashi, Y. *et al.* Mesp2 initiates somite segmentation through the Notch signalling pathway. *Nature Genet.* **25,** 390–396 (2000).
14. Bruckner, K., Perez, L., Clausen, H. & Cohen, S. Glycosyltransferase activity of Fringe modulates Notch-Delta interactions. *Nature* **406,** 411–415 (2000).
15. Moloney, D. J. *et al.* Fringe is a glycosyltransferase that modifies Notch. *Nature* **406,** 369–375 (2000).
16. Dale, J. K. *et al.* Periodic notch inhibition by lunatic fringe underlies the chick segmentation clock. *Nature* **421,** 275–278 (2003).
17. Bessho, Y., Hirata, H., Masamizu, Y. & Kageyama, R. Periodic repression by the bHLH factor Hes7 is an essential mechanism for the somite segmentation clock. *Genes Dev.* **17,** 1451–1456 (2003).
18. Hrabe de Angelis, M., McIntyre, J. II & Gossler, A. Maintenance of somite borders in mice requires the Delta homologue Dll1. *Nature* **386,** 717–721 (1997).
19. Evrard, Y. A., Lun, Y., Aulehla, A., Gan, L. & Johnson, R. L. lunatic fringe is an essential mediator of somite segmentation and patterning. *Nature* **394,** 377–381 (1998).
20. Cole, S. E., Levorse, J. M., Tilghman, S. M. & Vogt, T. F. Clock regulatory elements control cyclic expression of Lunatic fringe during somitogenesis. *Dev. Cell* **3,** 75–84 (2002).
21. Morales, A. V., Yasuda, Y. & Ish-Horowicz, D. Periodic Lunatic fringe expression is controlled during segmentation by a cyclic transcriptional enhancer responsive to notch signaling. *Dev. Cell* **3,** 63–74 (2002).
22. Bessho, Y. *et al.* Dynamic expression and essential functions of Hes7 in somite segmentation. *Genes Dev.* **15,** 2642–2647 (2001).
23. Sawada, A. *et al.* Fgf/MAPK signalling is a crucial positional cue in somite boundary formation. *Development* **128,** 4873–4880 (2001).
24. Dubrulle, J., McGrew, M. J. & Pourquie, O. FGF signaling controls somite boundary position and regulates segmentation clock control of spatiotemporal Hox gene activation. *Cell* **106,** 219–232 (2001).
25. Panin, V. M., Papayannopoulos, V., Wilson, R. & Irvine, K. D. Fringe modulates Notch–ligand interactions. *Nature* **387,** 908–912 (1997).
26. Haines, N. & Irvine, K. D. Glycosylation regulates Notch signalling. *Nature Rev. Mol. Cell Biol.* **4,** 786–797 (2003).
27. Hicks, C. *et al.* Fringe differentially modulates Jagged1 and Delta1 signalling through Notch1 and Notch2. *Nature Cell Biol.* **2,** 515–520 (2000).
28. Moreno, T. A. & Kintner, C. Regulation of segmental patterning by retinoic acid signaling during *Xenopus* somitogenesis. *Dev. Cell* **6,** 205–218 (2004).
29. Sakai, Y. *et al.* The retinoic acid-inactivating enzyme CYP26 is essential for establishing an uneven distribution of retinoic acid along the anterio-posterior axis within the mouse embryo. *Genes Dev.* **15,** 213–225 (2001).
30. Nagai, T. *et al.* A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nature Biotechnol.* **20,** 87–90 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Kitajima and E. Ikeno for their assistance in generating the *Mesp2–venus* knock-in mouse; M. Ikumi and Y. Takahashi for technical assistance and maintaining the mice used in this study; A. Gossler and R. Johnson for providing the Dll1and lunatic fringe knockout mice; and H. Hamada for *CYP26a*-null embryos. We also thank H. Takeda for critical reading of this manuscript. We are grateful to H. Tanaka for preparing purified Mesp2 protein. This work was supported by Grants-in-Aid for Science Research on Priority Areas (B), the Organized Research Combination System and National BioResource Project of the Ministry of Education, Culture, Sports, Science and Technology, Japan.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.S. (ysaga@lab.nig.ac.jp).

© 2005 Nature Publishing Group

LETTERS

The *Ter* mutation in the dead end gene causes germ cell loss and testicular germ cell tumours

Kirsten K. Youngren[1], Douglas Coveney[3], Xiaoning Peng[4], Chitralekha Bhattacharya[4], Laura S. Schmidt[5], Michael L. Nickerson[6], Bruce T. Lamb[1], Jian Min Deng[4], Richard R. Behringer[4], Blanche Capel[3], Edward M. Rubin[7], Joseph H. Nadeau[1,2] & Angabin Matin[4]

In mice, the *Ter* mutation causes primordial germ cell (PGC) loss in all genetic backgrounds[1]. *Ter* is also a potent modifier of spontaneous testicular germ cell tumour (TGCT) susceptibility in the 129 family of inbred strains, and markedly increases TGCT incidence in 129-*Ter*/*Ter* males[2–4]. In 129-*Ter*/*Ter* mice, some of the remaining PGCs transform into undifferentiated pluripotent embryonal carcinoma cells[2–6], and after birth differentiate into various cells and tissues that compose TGCTs. Here, we report the positional cloning of *Ter*, revealing a point mutation that introduces a termination codon in the mouse orthologue (*Dnd1*) of the zebrafish *dead end* (*dnd*) gene. PGC deficiency is corrected both with bacterial artificial chromosomes that contain *Dnd1* and with a *Dnd1*-encoding transgene. *Dnd1* is expressed in fetal gonads during the critical period when TGCTs originate. DND1 has an RNA recognition motif and is most similar to the apobec complementation factor, a component of the cytidine to uridine RNA-editing complex. These results suggest that *Ter* may adversely affect essential aspects of RNA biology during PGC development. DND1 is the first protein known to have an RNA recognition motif directly implicated as a heritable cause of spontaneous tumorigenesis. TGCT development in the 129-*Ter* mouse strain models paediatric TGCT in humans. This work will have important implications for our understanding of the genetic control of TGCT pathogenesis and PGC biology.

Ter was mapped to mouse chromosome 18 near *Fgf1* by following inheritance of genetic markers and two phenotypes: tumour incidence and PGC deficiency[7,8] (Fig. 1a–c). Testis weight was used as a surrogate quantitative phenotype for *Ter* in high-resolution interspecific and intersubspecific backcrosses and intercrosses (Supplementary Fig. 1). A total of 2,753 meioses were examined from crosses between the B6.129-*Ter* congenic strain and A/J, C57BL/6J, C3H/HeJ and MOLF/Ei strains to avoid recombination 'hot'- and 'cold'-spots that may exist in the genome of particular strains. Genotyping of F_2 progeny with microsatellite markers narrowed the critical interval to 0.14 cM between *273* and *173RB* (Fig. 2a). Four overlapping bacterial artificial chromosomes (BACs) spanned the *Ter* locus: RG-MBAC_173P21, RG-MBAC_282N17, RG-MBAC_273D11 and RG-MBAC_284F9, derived from the 129 strain.

We used BAC complementation tests to reduce further the size of the critical interval. Transgenic mice were generated with each of the four overlapping BACs. They were crossed to B6.129-*Ter* congenic mice to eventually generate *Ter*/*Ter* mice carrying individual transgenes. Mice with BACs 282N17 and 284F9, but not BACs 273D11 and 173P21, rescued the germ-cell-deficient phenotype (Supplementary Fig. 1), fully or partially restoring fertility in *Ter*/*Ter* mice, as assessed histologically (Fig. 2c, d) and by breeding performance (not shown), thus demonstrating that the overlapping region between 282N17 and 284F9 harbours *Ter*.

Four genes are present within the overlapping region of BACs

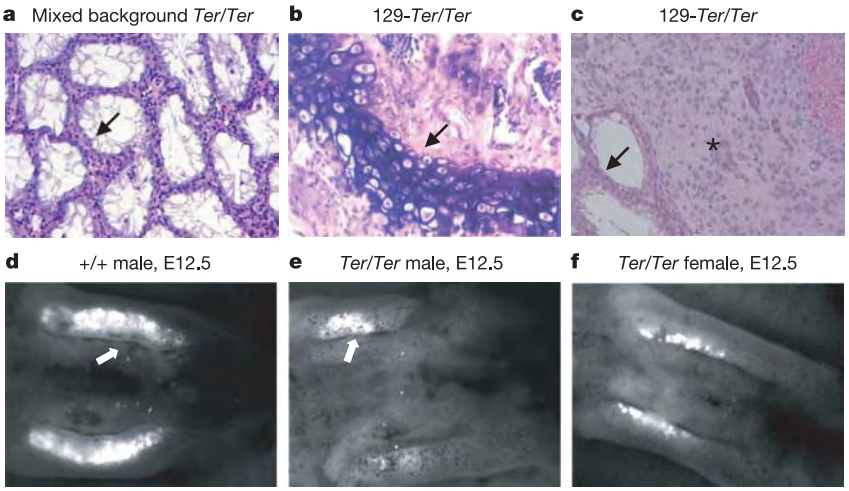


Figure 1 | Gonadal phenotypes of *Ter*/*Ter* males. **a**, Testis of a 4-week-old *Ter*/*Ter* male showing lack of germ cells (arrow points to Sertoli cell only phenotype) in the seminiferous tubules. **b**, **c**, Histological sections through TGCTs from 4-week-old 129-*Ter*/*Ter* male mice. Tissue types observed include cartilage (arrow in **b**) and neuroepithelia (asterisk in **c**). The arrow in **c** indicates germ-cell-deficient seminiferous tubules adjacent to neuroepithelial cells of the tumour. **d–f**, GFP-tagged PGCs. Genital ridges of E12.5 embryos showing GFP expression in PGCs of a 129-+/+;*Oct4*–GFP male (**d**), 129-*Ter*/*Ter*;*Oct4*–GFP male (**e**) and 129-*Ter*/*Ter*;*Oct4*–GFP female (**f**). The arrow indicates one of the genital ridges of the dissected embryo.

[1]Department of Genetics and [2]Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, Ohio 44106, USA. [3]Department of Cell Biology, Duke University, Durham, North Carolina 27710, USA. [4]Department of Molecular Genetics, University of Texas, MD Anderson Cancer Center, Houston, Texas 77030, USA. [5]Basic Research Program, SAIC-Frederick, Inc., and [6]Laboratory of Immunobiology, NCI Frederick, Frederick, Maryland 21702, USA. [7]Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA.

© 2005 Nature Publishing Group

282N17 and 284F9 (Fig. 2b). Comparison of the sequences of all the exons of these four genes from 129-*Ter*/*Ter* and 129-+/+ mice revealed a single base change (C to T) in the coding region of the *Dnd1* gene (Fig. 3a, b), which is the orthologue of zebrafish *dnd*[9]. This substitution introduces a stop codon (R178X) in the coding regions of both isoforms of mouse *Dnd1*-encoded protein (DND1) at amino acids 178 and 190, respectively (Fig. 3a; accession numbers BC034897 and AY321066, NCBI). The mutation was found only in mice carrying *Ter*, but not in 129 substrains (129T2/SvEmsJ, 129S1/SvImJ, 129X1/SvJ) or other inbred strains (A/J, C57BL/6J, C3H/HeJ and MOLF/Ei). The base change introduces a new *Dde*1 restriction enzyme site within the *Dnd1* sequence, enabling mice with the *Ter* allele to be distinguished from wild-type mice (Supplementary Fig. 1).

To confirm that *Ter* is due to mutation in *Dnd1*, we used a 3.9-kilobase (kb) genomic DNA fragment derived from BAC 284F9 that encodes only the *Dnd1* gene plus 860 base pairs (bp) and 510 bp flanking 5′ and 3′ ends, respectively, as a transgene (*Tg(Dnd1)1Matn*) to rescue the *Ter*/*Ter* phenotype. *Ter*/*Ter* males positive for *Tg(Dnd1)1Matn* showed partial rescue of the germ-cell-deficient phenotype (Fig. 2e). Although many seminiferous tubules remained germ-cell-deficient, some contained immature and mature sperm, which is unprecedented in *Ter*/*Ter* males.

Double-stranded RNAs, which can result from expression of overlapping genes on opposite strands, can lead to loss of messenger RNA or translational silencing[10,11]. *Dnd1* and *Wdr55* (2410080P20Rik) are encoded on opposite strands of DNA and overlap by 35 bp in the 3′ ends of their last exon (Fig. 2b). To test whether the mutation impairs function of *Wdr55* and indirectly leads to the *Ter* phenotype, we created mice with targeted deletion of *Wdr55* (*wdr55*^{tm1Matn}) (Supplementary Fig. 2). Mice homozygous for *wdr55*^{tm1Matn} are embryonic lethal and die before embryonic day (E)9.5. In contrast, *Ter*/*Ter* mice are viable but sterile. Moreover, double heterozygote males (+ *wdr55*^{tm1Matn}/*Ter* +) were not germ-cell-deficient and had gonad/body mass ratios ranging from 3.3 to 4.1, which is typical of wild-type but not *Ter*/*Ter* mice. Thus, *Ter* is not an allele of *Wdr55*. Therefore, functional complementation of *Tg(Dnd1)1Matn* transgenic mice together with normal testes in double heterozygotes demonstrates that the mutation in *Dnd1* directly causes the *Ter* phenotype.

Expression of *Dnd1*^{Ter} was characterized by northern and western blotting analyses. Northern analysis revealed that transcripts of both *Dnd1* isoforms are expressed in adult testes but only one isoform in adult heart (Fig. 3c). *Dnd1* transcripts were significantly decreased in tumours from 129-*Ter*/*Ter* males (Fig. 3d). Transcript levels of the other genes in the critical interval (*Wdr55*, *Ik* and *Ndufa2*) were also reduced to varying extents in tumour tissue, suggesting that these genes also function in normal adult testes. The two DND1 protein isoforms differ at the first 33 and 45 amino acids of their respective amino terminus. A polyclonal antibody was generated against an 18-amino-acid peptide of one isoform of DND1 (indicated by A in Fig. 3a). Western blotting of lysates from normal testes revealed a protein of 38 kDa, which is nearly the expected size (37.5 kDa) of the DND1 protein (Supplementary Fig. 1). This antibody also recognized a 62.5-kDa glutathione *S*-transferase (GST)–DND1 fusion protein. DND1 was not detected in protein lysates from germ-cell-deficient testes of three different B6.129-*Ter*/*Ter* males or from tumours of four different 129-*Ter*/*Ter* males (Fig. 3e). Thus, the various cell types of the tumours do not retain DND1 expression. However, a truncated DND1 protein was also not detected in *Ter*/*Ter* tumours. This rules out the possibility that tumour development is due to a dominant-negative-acting, oncogenic truncated DND1 protein. We note that the antibody we used detects only one isoform of DND1, and cannot rule out the possibility that the other isoform is

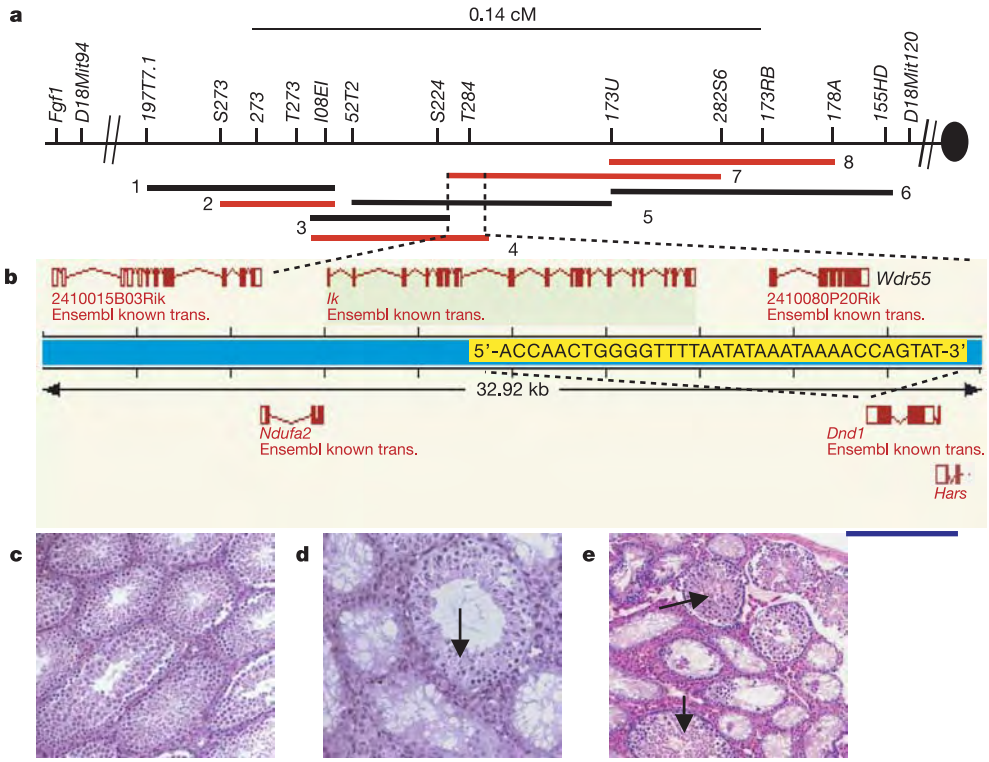


Figure 2 | Positional cloning of *Ter*. **a**, Physical map of the *Ter* locus. The 129-derived (RG-MBAC) BAC contig: 1 = 197J23, 2 = 173P21, 3 = 224K2, 4 = 284F9, 5 = 52N10, 6 = 368A6, 7 = 282N17, 8 = 273D11. **b**, Ensembl gene map of the overlapping region of BACs 284F9 and 282N17. Overlapping bases of the last exons of *Wdr55* and *Dnd1* are marked in yellow. The blue line indicates the DNA fragment used in *Tg(Dnd1)1Matn*. **c**, **d**, Testes histology of a *Ter*/*Ter* mouse with BAC 284F9 (**c**) and a sibling also with the transgene (**d**) where rescue occurs in a subset of seminiferous tubules (arrow). **e**, Histology of *Ter*/*Ter* testes with *Tg(Dnd1)1Matn* (the arrow shows a normal seminiferous tubule).

© 2005 Nature Publishing Group

still expressed. However, that seems to be unlikely because the mutation introduces a stop codon in both isoforms of DND1.

129-*Ter*/+ mice have a modestly elevated TGCT incidence, with 17% of the males having a unilateral tumour[2]. To test how *Dnd1* is involved in tumour development in heterozygous males, we obtained testicular tumours and contralateral normal testes from two 129-*Ter*/+ mice, and bisected each of these tissues to examine *Dnd1* protein and RNA expression. DND1 was not detected in tumour protein lysates (Fig. 3f), although the contralateral normal testes of the 129-*Ter*/+ mice expressed DND1. To determine the mechanism of protein loss, we examined RNA from the 129-*Ter*/+ tumour samples. Because the tumour sizes were typically small, we used polymerase chain reaction with reverse transcription (RT–PCR) to amplify the full-length *Dnd1* complementary DNA, and detected both the normal wild-type and mutated transcripts from tumour tissues via sequencing. We then developed an RT–PCR-based assay (see Methods) to distinguish visually mRNA from the normal (+) and the $Dnd1^{Ter}$ allele (Fig. 3g, left). Using this method, the wild-type transcript was observed in tumours as well as in normal testes (Fig. 3g, right). Because RT–PCR is a sensitive technique we were also able to amplify the $Dnd1^{Ter}$ transcript (Fig. 3g, labelled m) from the *Ter*/+ RNA. However, the $Dnd1^{Ter}$ transcript band is weaker compared with the wild-type allele. Taking into consideration both the western (Fig. 3f) and RT–PCR data (Fig. 3g, left), it indicates that inactivation of DND1 expression from the wild-type allele of 129-*Ter*/+ (a second 'hit') has resulted in tumour development. In the case of the two samples examined, the second hit is probably a post-transcriptional mechanism such as translational silencing[12,13].

The defect that underlies PGC deficiency and TGCT susceptibility in *Ter* males occurs during embryogenesis[1]. We therefore examined normal embryonic expression of *Dnd1* and detected it at earlier stages than reported previously[9]. Whole-mount *in situ* hybridization using a *Dnd1*-specific RNA probe detected expression in the embryo and allantoic bud at E7.5, neuroectoderm at E8.5, and widespread expression in E9.5 embryo including the neural tube, head mesenchyme, first branchial arch and the hindgut, through which primordial germ cells are migrating (Supplementary Fig. 3). *Dnd1* expression was also detected in the XY and XX genital ridges of E11.5 embryos (Fig. 4a, b). *Dnd1* expression was upregulated in the testis cords of the XY gonad between E12.5 and E14.5 (Figs 4c, e), but downregulated between E12.5 and E14.5 in XX gonads (Fig. 4d, f). Downregulation of *Dnd1* in the XX gonad occurred progressively from anterior to posterior, similar to the wave of meiotic entry of XX germ cells[14,15]. The contrasting patterns of *Dnd1* expression in XX versus XY gonads may account for the differential susceptibility to teratocarcinogenesis in female versus male 129-*Ter*/*Ter* mice.

To visualize PGCs in 129-*Ter*/*Ter* embryos, we introduced a green

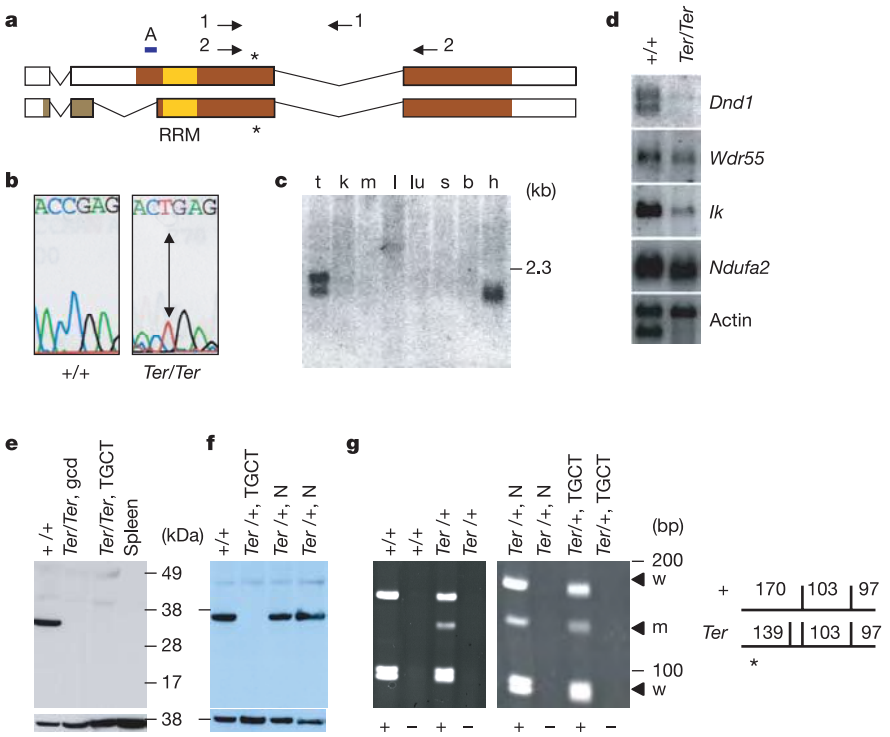


Figure 3 | Expression of *Dnd1* in normal tissues and TGCTs. **a**, Comparison of DND1 isoforms with accession numbers BC034897 (top) and AY321066 (bottom). Coloured regions indicate protein-coding regions; yellow box marks an RNA recognition motif. An asterisk indicates amino acids 178 and 190, of the two isoforms, which are mutated (R178X) in *Ter*. Arrows indicate primers 4.13a (1) and A25 (2). A indicates the region used as peptide epitope to generate anti-DND1 antibody. **b**, C to T mutation in *Ter*/*Ter* introduces a stop codon. **c**, Mouse tissue northern blot for *Dnd1*. mRNAs are from testes (t), kidney (k), muscle (m), liver (l), lung (lu), spleen (s), brain (b) and heart (h). **d**, Northern blot of mRNA from 129-+/+ testes and 129-*Ter*/*Ter* tumours. **e**, Western blot with anti-DND1 antibody of tissue lysates from normal (+/+) testes of a 129-+/+ mouse, germ-cell-deficient (gcd) testes of a B6.129-*Ter*/*Ter* congenic strain, bilateral TGCTs from a 129-*Ter*/*Ter* mouse, and spleen of a 129-+/+ mouse. The bottom panel shows control for loading using anti-β-actin. **f**, Western blot with anti-DND1 of normal testes of a 129-+/+ mouse, tumour in the testes of a 129-*Ter*/+ (*Ter*/+, TGCT) mouse, the normal contralateral testes (*Ter*/+, N) and normal testes of another 129-*Ter*/+ mouse (*Ter*/+, N). The bottom panel shows the same blot re-hybridized with mouse anti-β-actin. **g**, RT–PCR of RNA from normal testes of 129-+/+ and 129-*Ter*/+ mice using the A25 primer (2 in panel **a**) followed by *Dde*1 digestion (left panel). This produces fragments of 170, 103 and 97 bp from the wild-type allele and 139, 103, 97 and 31 bp (not seen on gel) from the *Ter* allele. For controls, PCR reactions were performed on samples that included (+) or did not include (−) Superscript II during RT, as indicated below the panels. The right panel shows total RNA from normal and TGCT-bearing testes of the same 129-*Ter*/+ mouse, amplified by RT–PCR followed by *Dde*1 digestion and electrophoresis. The arrows indicate the wild-type (w) and *Ter* (m) allele.

© 2005 Nature Publishing Group

fluorescent protein (GFP) transgene driven by the PGC-specific Oct4 promoter[16] (*GOF-1/ΔPE/EGFP*) in the 129-*Ter* strain. At E12.5, the number of GFP-labelled PGCs was significantly reduced in the urogenital ridges of 129-*Ter*/*Ter* compared with 129-+/+ embryos (Fig. 1d–f). However, the few surviving PGCs in *Ter*/*Ter* embryos had successfully migrated to the genital ridges[1], and tumours eventually occur in the testes of *Ter*/*Ter* mice. Therefore, *Dnd1* is not essential for PGC migration in the mouse. Loss of PGCs in 129-*Ter*/*Ter* embryos precedes development of embryonal carcinoma cells and TGCTs. Inactivation of *Dnd1* expression in PGCs during embryogenesis in 129-*Ter*/*Ter* mice is therefore implicated as the causal event that drives PGCs to exit the germ line and transform to embryonal carcinoma cells and TGCTs.

The mouse DND1 protein shares sequence similarity with proteins with RNA recognition motifs, with highest similarity to mouse apobec complementation factor (Acf, XP_129159; 34% overall amino acid identity and 48% identity in the RNA recognition motif) (Supplementary Fig. 4). *Acf* encodes the RNA-binding subunit of the RNA-editing enzyme complex (editosome) that converts specific cytidines to uridines in the apolipoprotein B transcript and other RNAs[17]. Several RNA-binding proteins are anomalously expressed in certain cancers[18–23], but it is unclear whether these are causes or consequences of tumorigenesis. DND1 may be an RNA- or DNA-binding subunit of other apobec-like proteins[24] involved in nucleic-acid editing, and the *Ter* mutation raises the possibility that TGCT susceptibility is a consequence of aberrant nucleic acid editing.

Various aspects of RNA biology have central roles in the development of the PGC lineage[25–27]. We show that functional inactivation of the mouse *Dnd1* gene results in severe germ cell deficiency and increased TGCT susceptibility. Our findings implicate a causal role of RNA biology in TGCT susceptibility, as DND1 has an RRM motif[28] and is most similar to Acf, which is involved in RNA editing. Extrapolating from the findings in *Ter* mice, although germ cell tumours present clinically in infants and young adults, it is apparent that genetic and environmental influences during embryogenesis increase the susceptibility of PGCs to tumorigenesis. 129-*Ter* mice are a valuable model system to study the initiating events of tumorigenesis, because the cell of origin of the tumours and the time of onset is defined, and now one of the defects (the *Ter* mutation in *Dnd1*) that predisposes to PGC transformation has been identified.

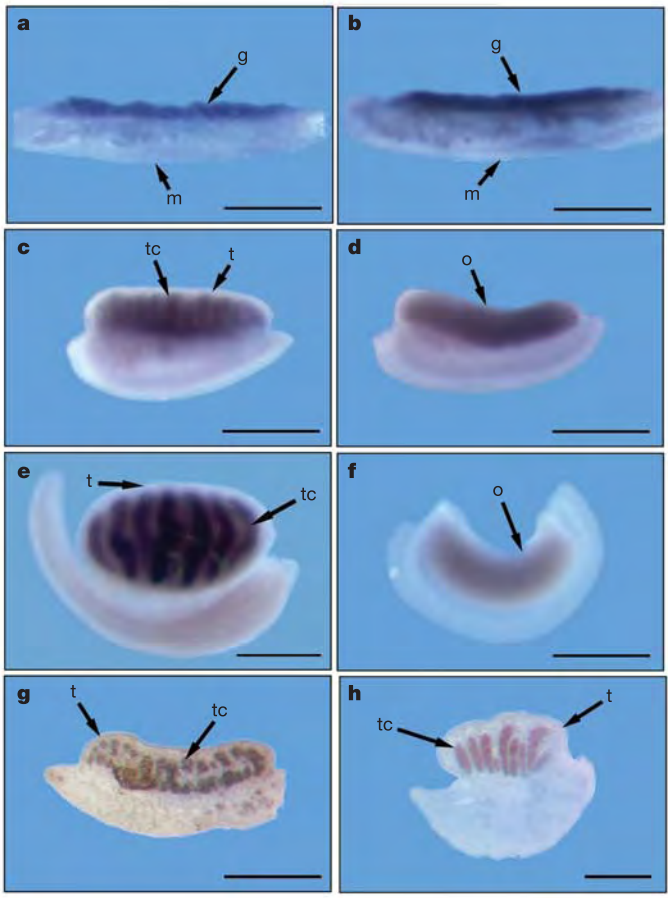


Figure 4 | Expression of *Dnd1* in embryonic gonads. a, **b**, Whole-mount *in situ* hybridization of E11.5 XY (**a**) and E11.5 XX (**b**) genital ridge (g) and mesonephros (m), in which *Dnd1* expression is localized to the gonad. **c**, **d**, Whole-mount *in situ* hybridization of an E12.5 XY (**c**) gonad (t), in which *Dnd1* is localized to testis cords (tc), whereas *Dnd1* expression in an E12.5 XX (**d**) gonad (o) has a broad distribution throughout the gonad. **e**, **f**, Whole-mount *in situ* hybridization of an E14.5 XY (**f**) gonad, in which *Dnd1* is localized to testis cords, in contrast to E14.5 XX (**f**) gonads, in which *Dnd1* expression is reduced. **g** and **h** show section *in situ* hybridization of an E12.5 XY and E13.5 XY gonad, respectively. *Dnd1* is specific to developing testicular cords in the E12.5 and E13.5 XY gonads. Scale bars represent 500 μm.

METHODS

Nomenclature. The orthologue of *Ter* in zebrafish is the *dead end* (*dnd*) gene[9]. On the basis of orthology with *dnd*, we call the mouse gene dead end, with gene and allele symbol *Dnd1* and *Dnd1*^{Ter}, respectively. Mouse *Dnd1* accession numbers are BC034897 and AY321066 (NCBI). The gene 2410080P20Rik (NM_026464) on the opposite strand and partially overlapping with *Dnd1* was named *Wdr55*. Accession numbers for *Ik* and *Ndufa2* are NM_011879 and NM_010885, respectively.

Generation of genetic and physical maps. Polymorphic microsatellites were subcloned from the BACs using colony hybridization and mapped to chromosome 18 (primers to amplify the polymorphic markers are listed in Supplementary Fig. 4). Accession numbers of the BACs are: AC027276 for RG-MBAC_173P21, AC027278 for RG-MBAC_284F9, AC087795 for RP23-181H5 (which is the C57BL/6J equivalent of RG-MBAC_282N17) and AC087772 for RP23-326L17 (which is the C57BL/6J equivalent of RG-MBAC_2849F). BAC RG-MBAC_273D11 was not sequenced.

Creation of BAC and *Tg(Dnd1)1Am* transgenics. BACs were purified, linearized and microinjected into FVB fertilized one-cell embryos[29]. PCR analysis to test for the presence of the BAC transgenes was carried out using the primers BAC-51-F 5′-CTTAACTATGCGGCATCAGAGC-3′ and BAC-201-R 5′-GCCAGCTGGCGTAATAGCGAAG-3′, which amplified a 173-bp fragment from the vector-insert junction. Mice carrying the 129-derived chromosomal segment spanning *D18Mit232*, *D18Mit94* and *D18Mit17* (*Ter*/+) and mice positive for BAC transgenes were intercrossed. Genotyping for the 129-derived chromosomal segment identified *Ter*/*Ter* mice bearing individual BAC transgenes. To generate *Tg(Dnd1)1Matn*, BAC 284F9 was digested with *Bgl*II and *Nhe*1 and subcloned into pBlueScript II KS+. The 3.9-kb fragment encoding *Dnd1* was identified by hybridization with ^{32}P-labelled PCR product of primers 4.13a-F and 4.13a-R (Supplementary Methods). The 3.9-kb fragment was purified, linearized and microinjected into FVB-fertilized one-cell embryos. Primers designed against the vector-insert boundary, A16.1-F 5′-GTCATCC TTGCACGCTGCGCACG-3′ and A16.1-R 5′-GCTTGATATCGAATTCCTG-3′, were used for PCR genotyping for the presence of *Tg(Dnd1)1Matn*. Founder mice were crossed to the B6.129-*Ter* congenic strain. Mice carrying the 129-derived chromosomal segment spanning *D18Mit232*, *D18Mit94* and *D18Mit17* (*Ter*/+) and positive for the *Tg(Dnd1)1Matn* transgene were intercrossed to generate *Ter*/*Ter* mice bearing the *Tg(Dnd1)1Matn* transgene. Testis weight measurements were taken for male progeny and stained sections were examined histologically.

Gonad/body weight ratios and histology. Adult mice (4 weeks of age or older) were killed and the body weight and testis weight recorded. The gonad/body weight ratio was calculated as (gonad weight/body weight) × 1,000. Testes or tumours were preserved in 10% phosphate-buffered formalin for at least 48 h. Sections (5 μm) were stained with haematoxylin and eosin.

***In situ* hybridization.** Embryos and gonads were fixed in 4% PFA for 12 h at 4 °C, washed in 0.1% Tween 20/PBS, dehydrated in 100% methanol and stored at 20 °C until used. Full-length cDNA of *Dnd1* (image clone ID: 3470697) was used as probe. *In situ* hybridizations on mouse embryos were performed as described previously[30].

© 2005 Nature Publishing Group

Generation of anti-DND1 antibody and GST–DND1. Rabbit polyclonal antibody was generated against the DND1 peptide spanning amino acids 16–33. GST–DND1 fusion protein was made by cloning *Dnd1* cDNA in-frame into pGEX-2TK (Amersham Biosciences). GST–DND1 fusion protein was induced by isopropyl-β-D-thiogalactoside and affinity purified using glutathione Sepharose 4B.

Assays to distinguish the *Dnd* alleles. To distinguish the mRNA transcripts from the wild-type and *Ter* alleles of *Dnd1*, total RNA from 129-+/+ and 129-*Ter*/+ testes and tumours were treated with DNase1 before converting to cDNA using Superscript II Rnase H Reverse Transcriptase (Invitrogen). PCR was performed using the primers A25-F 5′-CGAGCTGACGGTGGACGGGCTGCC-3′ and A25-R 5′-CTGCCTAGCTTCATCCGTTGAC-3′ (primers 2, Fig. 3a), which were designed to span exons 2 and 3 of *Dnd1* to yield a 396-bp product. The RT–PCR products were digested with *Dde*1 before electrophoresis on a 7% acrylamide gel. *Dde*1 digestion of RT–PCR product produces fragments of 170, 103 and 97 bp from the wild-type allele and 139, 103, 97 and 31 bp from the *Ter* allele. The 31-bp fragment was not observed on the gel (Fig. 3g). Controls to exclude amplification of contaminating DNA were performed on parallel RNA samples by leaving out Superscript II during the reverse transcription reaction (indicated in the lanes marked −). The primers Dnd13-F (5′-CAGGAGCCAG AAAGGTATCAATA-3′) and Dnd13-R (5′-CTTAACCCATTTAGGGTACCTGT-3′) were used to amplify a 1,059-bp *Dnd1* transcript by RT–PCR for sequencing.

Received 9 December 2004; accepted 24 March 2005.

1. Sakurai, T., Iguchi, T., Moriwaki, K. & Noguchi, M. The *ter* mutation first causes primordial germ cell deficiency in *ter/ter* mouse embryos at 8 days of gestation. *Dev. Growth Differ.* **37,** 293–302 (1995).
2. Noguchi, T. & Noguchi, M. A recessive mutation (*ter*) causing germ cell deficiency and a high incidence of congenital testicular teratomas in 129/Sv-*ter* mice. *J. Natl. Cancer Inst.* **75,** 385–392 (1985).
3. Stevens, L. C. A new inbred subline of mice (129-terSv) with a high incidence of spontaneous congenital testicular teratomas. *J. Natl Cancer Inst.* **50,** 235–242 (1973).
4. Stevens, L. C. & Mackensen, J. A. Genetic and environmental influences on teratocarcinogenesis in mice. *J. Natl Cancer Inst.* **27,** 443–453 (1961).
5. Stevens, L. C. Origin of testicular teratomas from primordial germ cells in mice. *J. Natl Cancer Inst.* **38,** 549–552 (1967).
6. Donovan, P. J. & de Miguel, M. P. Turning germ cells into stem cells. *Curr. Opin. Genet. Dev.* **13,** 463–471 (2003).
7. Asada, Y., Varnum, D. S., Frankel, W. N. & Nadeau, J. H. A mutation in the *Ter* gene causing increased susceptibility to testicular teratomas maps to mouse chromosome 18. *Nature Genet.* **6,** 363–368 (1994).
8. Sakurai, T., Katoh, H., Moriwaki, K., Noguchi, T. & Noguchi, M. The *ter* primordial germ cell deficiency mutation maps near *Grl-1* on mouse chromosome 18. *Mamm. Genome* **5,** 333–336 (1994).
9. Weidinger, G. *et al. dead end*, a novel vertebrate germ plasm component, is required for zebrafish primordial germ cell migration and survival. *Curr. Biol.* **13,** 1429–1434 (2003).
10. Mello, C. C. & Conte, D. Revealing the world of RNA interference. *Nature* **431,** 338–342 (2004).
11. Meister, G. & Tuschl, T. Mechanisms of gene silencing by double-stranded RNA. *Nature* **431,** 343–349 (2004).
12. Shao, J., Sheng, H., Inoue, H., Morrow, J. D. & DuBois, R. N. Regulation of constitutive cyclooxygenase-2 expression in colon carcinoma cells. *J. Biol. Chem.* **275,** 33951–33956 (2000).
13. Mukhopadhyay, D., Houchen, C. W., Kennedy, S., Dieckgraefe, B. K. & Anant, S. Coupled mRNA stabilization and translational silencing of cyclooxygenase-2 by a novel RNA binding protein, CUGBP2. *Mol. Cell* **11,** 113–126 (2003).
14. Yao, H. H., DiNapoli, L. & Capel, B. Meiotic germ cells antagonize mesonephric cell migration and testis cord formation in mouse gonads. *Development* **130,** 5895–5902 (2003).
15. Menke, D. B., Koubova, J. & Page, D. C. Sexual differentiation of germ cells in XX mouse gonads occurs in an anterior-to-posterior wave. *Dev. Biol.* **262,** 303–312 (2003).
16. Scholer, H. R., Dressler, G. R., Balling, R., Rohdewohld, H. & Gruss, P. Oct-4: a germline-specific transcription factor mapping to the mouse t-complex. *EMBO J.* **9,** 2185–2195 (1990).
17. Mehta, A., Kinter, M. T., Sherman, N. E. & Driscoll, D. M. Molecular cloning of apobec-1 complementation factor, a novel RNA-binding protein involved in the editing of apolipoprotein B mRNA. *Mol. Cell. Biol.* **20,** 1846–1854 (2000).
18. Ma, Z. *et al.* Fusion of two novel genes, RBM15 and MKL1, in the t(1;22)(p13;q13) of acute megakaryoblastic leukemia. *Nature Genet.* **28,** 220–221 (2001).
19. Barbouti, A. *et al.* A novel gene, MSI2, encoding a putative RNA-binding protein is recurrently rearranged at disease progression of chronic myeloid leukemia and forms a fusion gene with HOXA9 as a result of the cryptic t(7;17)(p15;q23). *Cancer Res.* **63,** 1202–1206 (2003).
20. Drabkin, H. A. *et al.* DEF-3 (g16/NY-LU-12), an RNA binding protein from the 3p21.3 homozygous deletion region in SCLC. *Oncogene* **18,** 2589–2597 (1999).
21. Ross, J., Lemm, I. & Berberet, B. Overexpression of an mRNA-binding protein in human colorectal cancer. *Oncogene* **20,** 6544–6550 (2001).
22. Jinawath, N., Furukawa, Y. & Nakamura, Y. Identification of NOL8, a nucleolar protein containing an RNA recognition motif (RRM), which is overexpressed in diffuse-type gastric cancer. *Cancer Sci.* **95,** 430–435 (2004).
23. Tsuei, D.-J. *et al.* RBMY, a male germ cell-specific RNA-binding protein, activated in human liver cancers and transforms rodent fibroblasts. *Oncogene* **23,** 5815–5822 (2004).
24. Wedekind, J. E., Dance, G. S. C., Sowden, M. P. & Smith, H. C. Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. *Trends Genet.* **19,** 207–216 (2003).
25. Martinho, R. G., Kunwar, P. S., Casanova, J. & Lehmann, R. A noncoding RNA is required for the repression of RNA polII-dependent transcription in primordial germ cells. *Curr. Biol.* **14,** 159–165 (2004).
26. Moore, F. L. *et al.* Human pumilio-2 is expressed in embryonic stem cells and germ cells and interacts with DAZ (Deleted in AZoospermia) and DAZ-like proteins. *Proc. Natl Acad. Sci. USA* **100,** 538–543 (2003).
27. Crittenden, S. L. *et al.* A conserved RNA-binding protein controls germline stem cells in *Caenorhabditis elegans*. *Nature* **417,** 660–663 (2002).
28. Burd, C. G. & Dreyfuss, G. Conserved structures and diversity of functions of RNA-binding proteins. *Science* **265,** 615–621 (1994).
29. Yang, X. W., Model, P. & Heintz, N. Homologous recombination based modification in *Escherichia coli* and germline transmission in transgenic mice of a bacterial artificial chromosome. *Nature Biotechnol.* **15,** 859–865 (1997).
30. Henrique, D. *et al.* A digoxigenin labeled RNA probe for Sox9 was detected using an alkaline phosphatase-conjugated anti-digoxigenin antibody. *Nature* **375,** 787–790 (1995).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. Scholer for the *GOF-1/ΔPE/EGFP* construct, A. Kong and W. Cosme-Blanco for technical help, G. Lozano for critical reading of the manuscript, and members of the Nadeau laboratory for suggestions. Services of the Trans-NIH Mouse Initiative were used for sequencing BACs encoding the *Ter* locus. This project was supported by NCI grants to A.M. and J.H.N. and with funds from the NCI and NIH to L.S.S. D.C. and B.C. are funded by a grant from NIH. Veterinary resources, DNA sequencing and Genetically Engineered Mouse Facility were supported by a Cancer Center Support (Core) Grant. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.M. (amatin@mdanderson.org) or J.H.N. (jhn4@case.edu).

© 2005 Nature Publishing Group

LETTERS

Non-equilibration of hydrostatic pressure in blebbing cells

Guillaume T. Charras[1], Justin C. Yarrow[1], Mike A. Horton[2], L. Mahadevan[1,3,4] & T. J. Mitchison[1]

Current models for protrusive motility in animal cells focus on cytoskeleton-based mechanisms, where localized protrusion is driven by local regulation of actin biochemistry[1–3]. In plants and fungi, protrusion is driven primarily by hydrostatic pressure[4–6]. For hydrostatic pressure to drive localized protrusion in animal cells[7,8], it would have to be locally regulated, but current models treating cytoplasm as an incompressible viscoelastic continuum[9] or viscous liquid[10] require that hydrostatic pressure equilibrates essentially instantaneously over the whole cell. Here, we use cell blebs as reporters of local pressure in the cytoplasm. When we locally perfuse blebbing cells with cortex-relaxing drugs to dissipate pressure on one side, blebbing continues on the untreated side, implying non-equilibration of pressure on scales of approximately 10 μm and 10 s. We can account for localization of pressure by considering the cytoplasm as a contractile, elastic network infiltrated by cytosol. Motion of the fluid relative to the network generates spatially heterogeneous transients in the pressure field, and can be described in the framework of poroelasticity[11,12].

Blebs are large, approximately spherical deformations of the cell surface that form and disappear on a timescale of tens of seconds. Although less studied than lamellipodial or filopodial protrusion, blebbing is a common phenomenon during apoptosis[13], cytokinesis[14,15] and cell movement[16,17]. Blebs are thought to be initiated by rupture of the plasma membrane from the underlying cytoskeleton (Supplementary Fig. 3), followed by inflation of the detached membrane by intracellular fluid flow[18,19]. Subsequent bleb retraction is driven by assembly and contraction of a cortex within the newly formed bleb. Blebbing requires non-uniform behaviour of the membrane and cortex that must reflect either a globally uniform hydrostatic pressure combined with local nucleation of membrane detachment from the cytoskeleton, or non-uniform pressure leading to local rupture of membrane–cytoskeleton attachments in regions of high pressure. To distinguish these possibilities, we studied the effect of local disruption of cortical contraction or integrity induced by drug perfusion over part of a cell, in a filamin-depleted melanoma cell line (M2) that blebs extensively and continuously[20].

We first confirmed that the cell volume stays approximately constant during blebbing (Supplementary Data, Supplementary Video 12 and Supplementary Fig. 4)[18], implying that blebbing is driven primarily by flow of fluid within the cell rather than water crossing the plasma membrane. We then confirmed that bleb inflation represents ballooning out of the plasma membrane when it detaches from the actin cortex by simultaneously imaging the cortex and the cell membrane (Supplementary Fig. 3). During bleb inflation, green fluorescent protein (GFP)-tagged actin was not enriched at the bleb surface compared to its interior, implying lack of a cytoskeleton in the bleb (Fig. 1a, $t = 0$ and 22 s). As inflation slowed, a cortical cytoskeleton assembled underneath the bleb membrane, evidenced by a rim of actin fluorescence (Fig. 1a, $t = 62$ s); simultaneously, the cortex at the base of the bleb disassembled (Supplementary Fig. 3). As the bleb retracted, its actin rim became more pronounced, and often wavy or buckled, suggesting that the cortex contracts as the bleb shrinks (Fig. 1a, $t = 72$ s)[19]. GFP–myosin II regulatory light chain (MRLC) was recruited to the newly forming cortex of the bleb simultaneous with actin (Fig. 1b; see also Supplementary Video 10)[13,21], consistent with myosin II driving contraction. Increasing extracellular osmolarity made blebs smaller, whereas decreasing it made them larger[22] (Fig. 1; see also Supplementary Video 1). Increasing membrane rigidity by crosslinking the glycocalix polysaccharides with wheat germ agglutinin (WGA)[23] inhibits blebbing (Supplementary Video 2). These observations, together with drug studies discussed below, support the following qualitative model for bleb dynamics[19]. Cortical acto-myosin contracts (Supplementary Videos 10 and 11), generating hydrostatic pressure that causes a patch of plasma membrane to tear free from its attachment to the cortical cytoskeleton. This patch of cytoskeleton-free membrane rapidly inflates as cytosol flows in, with its base enlarging by further tearing (Supplementary Fig. 3). Later, inflation slows and a mesh of actin and myosin II assembles in the bleb to form a contractile cortex attached to the plasma membrane (Fig. 1b). Finally, contraction of this cortical mesh causes the bleb to shrink, driving the extruded cytosol back into the cell body.

To find inhibitory small molecules, we screened a library of known biologically active compounds for rapid inhibition of blebbing. After characterizing their effects in bath treatment, selected drugs and osmolytes were applied locally to cells by injecting medium containing inhibitor via a micropipette into a laminar fluid flow[24,25]. The treated region was visualized by adding a fluorescent tracer to the pipette medium, whereas blebbing was observed using differential interference contrast (DIC) microscopy. During local perfusion, we typically bathed 20–33% of the cell area. We confirmed local application by visualizing local deposition of a fluorescent lectin (WGA–Alexa 488; Fig. 2a, $t = 430$ s). For each drug, local perfusion experiments were repeated until one of the following behaviours was observed reproducibly ($N \geq 5$): (1) local perfusion of drug inhibited blebbing locally; (2) local perfusion had both a local and global effect on blebbing; (3) local perfusion did not inhibit blebbing but whole-cell treatment did.

The effect of local perfusion could be classified into three categories. The first is treatments that locally inhibited blebbing but allowed the untreated part of the cell to bleb normally (Fig. 2; see also Supplementary Table 1 and Supplementary Videos 3–6). As expected, agents that increased membrane rigidity (WGA) or osmotic pressure (sucrose) had this effect (Fig. 2a, b). These agents locally increase the physical forces that oppose blebbing. Notably, drugs that block generation of contractile force also had local inhibitory effects, including inhibitors of myosin II (blebbistatin[21]) and ROCK1 (Y-27632 and 3-(4-pyridyl)indole[26]), a kinase that activates

[1]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA. [2]London Centre for Nanotechnology, University College London, London WC1E 6JF, UK. [3]Division of Engineering and Applied Sciences, and [4]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA.

© 2005 Nature Publishing Group

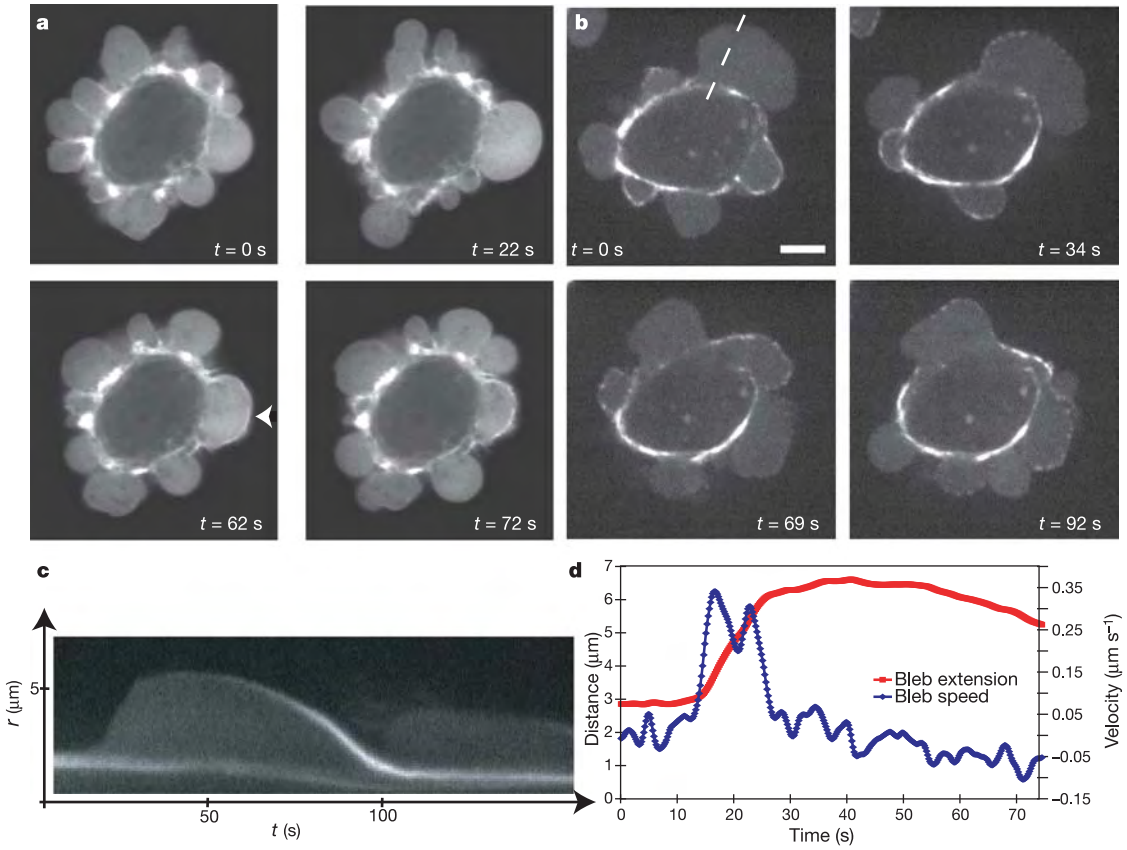


Figure 1 | Localization of GFP-actin and GFP-MRLC in blebs shows that expansion is a passive process and retraction an active process necessitating actin and MRLC. All images were acquired by confocal microscopy. Scale bar, 5 μm. **a**, GFP–actin-transfected cell. During bleb expansion, the bleb rim does not appear to be enriched in GFP–actin compared to the bleb interior ($t = 0$ s and $t = 22$ s). Once expansion has halted, an actin-rich rim forms at the bleb surface ($t = 62$ s, arrowhead) and retraction begins. As retraction proceeds, the actin rim becomes wavy, suggesting that the cortex contracts as the bleb shrinks ($t = 72$ s). **b**, GFP–MRLC-transfected cell. Just after expansion has stopped, the bleb surface does not appear to be enriched in GFP–MRLC ($t = 0$ s). MRLC then accumulates in discrete foci ($t = 34$ s) and retraction starts ($t = 69$ s). As retraction ends, MRLC forms a continuum along the bleb rim ($t = 92$ s). **c**, Kymograph of the expansion and retraction of the bleb in **b** (dashed line). As the bleb expands, there is no accumulation of MRLC at the bleb apex. As retraction proceeds, MRLC accumulates at the bleb. Bleb extension (r) is on the y axis. **d**, Graph of the bleb extension and bleb velocity as a function of time for the bleb shown in **c**. The maximal bleb extension is approximately 3.5 μm (maximal speed of expansion about 0.35 μm s^{-1}, speed of retraction about 0.1 μm s^{-1}).

myosin II (Fig. 2c, d). The second category is treatments that caused a combination of local and global effects, as seen with the F-actin-disrupting agents cytochalasin D and latrunculin B (Fig. 3a; see also Supplementary Video 7). Within seconds of local drug application, there was a global increase in bleb size (Fig. 3a, $t = 75$ s). Then, blebbing ceased in the treated area and existing blebs were not retracted, whereas blebbing continued in the untreated area at a slower rate than before drug application (Fig. 3a, $t = 315$ s). The third category is treatments that only had an effect when the whole cell was treated. Inhibitors of several other protein kinases had this effect (Fig. 3b, c; see also Supplementary Videos 8 and 9). For local perfusion to perturb the cytoplasm locally, membrane crossing must be fast compared with intracellular diffusion of the drug or drug–target complex[27]. Thus, failure to see a local effect of a drug cannot be interpreted in terms of local versus global action of its target.

The local effects of myosin-II-inhibiting drugs (Fig. 2c, d) show that the acto-myosin cortex acts locally to promote blebbing. This could be because contraction locally nucleates blebs, while hydrostatic pressure is uniform, or because pressure generated by contraction only acts locally to extrude blebs, implying that it does not globally equilibrate. Local inhibition of blebbing by actin-depolymerizing drugs (Fig. 3a) favours the latter hypothesis. If pressure equilibrated across the cell, the drug-treated side, where the cortex is softer, should swell preferentially, inhibiting blebbing on the untreated side. Because this does not happen, we conclude that hydrostatic pressure does not equilibrate across single M2 cells on time- and length-scales relevant to motility.

Current models of the cytoplasm[9,10] cannot account for spatio-temporally localized variations in hydrostatic pressure. We therefore propose a new description of the cytoplasm based on poro-elasticity[11,12]. We consider the cytoplasm to be composed of a porous, actively contractile, elastic network (cytoskeletal filaments, organelles and ribosomes), infiltrated with an interstitial fluid (that is, the cytosol, comprising water, ions and soluble proteins), similar to a fluid-filled sponge. Contraction of the acto-myosin cortex creates a compressive stress on the cytoskeletal network, leading to a spatially localized increase in hydrostatic pressure (Fig. 4a). In response, the cytosol flows out of the network, and if it finds a region where the membrane is weakly attached to the cortex; the resulting pressure can lead to membrane detachment and bleb inflation (Fig. 4b; see also Supplementary Fig. 3). With a poroelastic description of the cytoplasm, hydrostatic pressure does not instantaneously propagate through the network. Instead, it diffuses over a length $x \approx \sqrt{Dt}$ during a time t, with the diffusion constant $D = \frac{kK}{\phi}$ determined by an appropriately defined elastic bulk modulus K, the hydraulic permeability of the network k, and the local volume fraction of fluid ϕ (see Supplementary Information). The diffusion constant D dictates the time needed for the effect of a local contractile force to be felt in

© 2005 Nature Publishing Group

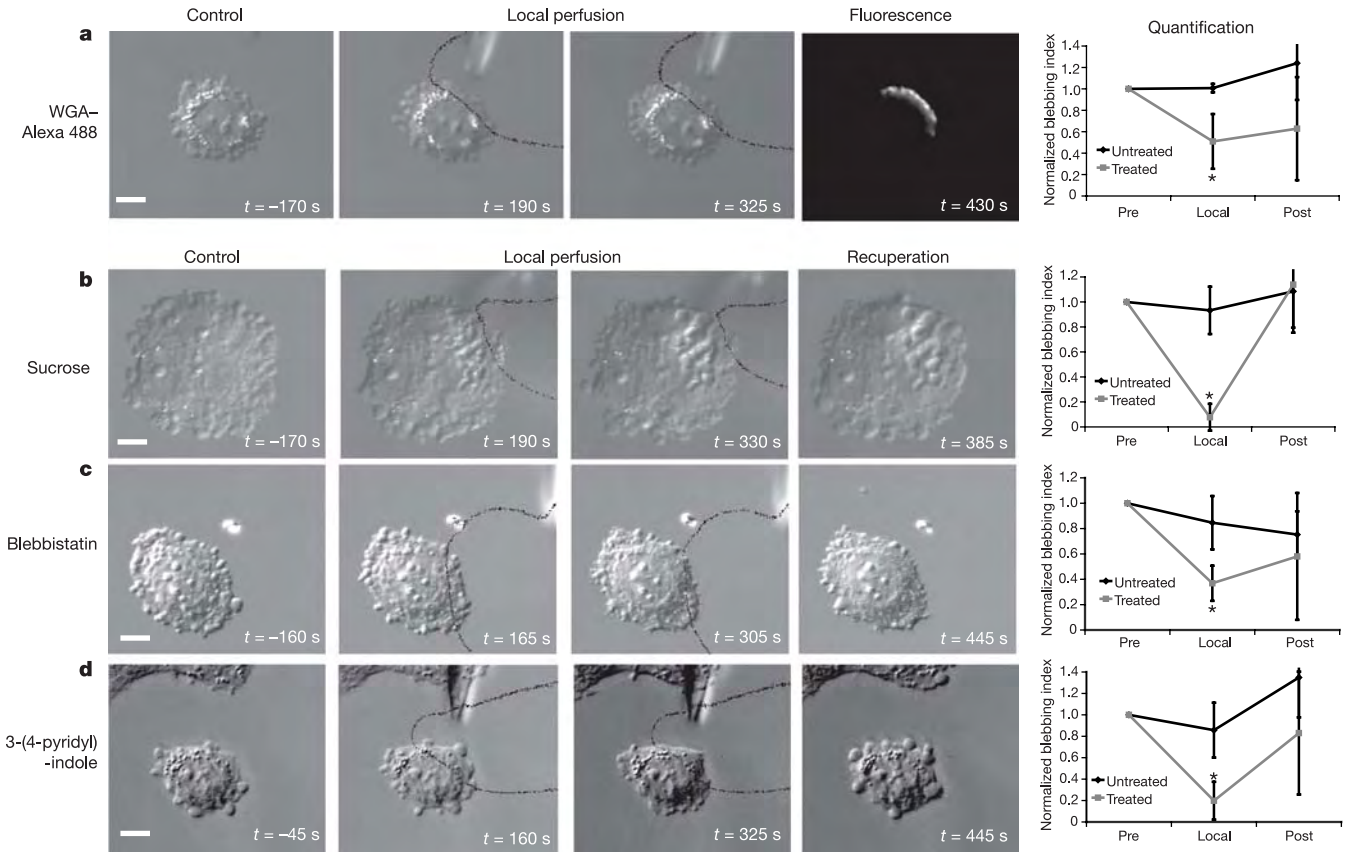


Figure 2 | Local perfusion can locally inhibit blebbing. In all cases, before application of treatment the cells blebbed profusely around their periphery.Within 2–3 min of application, blebbing ceases in the area of application but continues undisturbed in the rest of the cell. In some cases, when the local perfusion was halted ($t_{\mathrm{off}} = 350$ s in all cases), cells recovered and blebbed normally. **a**, Local perfusion of WGA–Alexa 488. WGA–Alexa 488 was only incorporated in the region of the cell that was exposed to flow from the micropipette. **b**, Local perfusion of 300 mM sucrose solution. Vacuoles form in the vicinity of the perfused region. **c**, Local inhibition of myosin II ATPase by local perfusion of blebbistatin. **d**, Local inhibition of ROCKI by 3-(4-pyridyl)indole. Black lines delineate the flow out of the micropipette. On each image, timing relative to local application of treatment is given. The normalized blebbing indices (right) show the evolution over time of blebbing in the region exposed to inhibitor and in the free region. Error bars show the standard deviation. Asterisks denote significant changes in the blebbing index when compared with the initial blebbing index. Scale bars, 10 μm.

other parts of the cell. Using a typical timescale for bleb inflation of $t \approx 10$ s and realistic values for the various parameters (Supplementary Information), we find $x \approx 15$–$30\,\mu$m. Hence, hydrostatic pressure can be strongly non-equilibrated in cells on a timescale of approximately 10 s and a length scale of about 10 μm—scales that are relevant to a variety of motile behaviours. Bleb formation reduces this length scale by allowing the fluid to flow, with less resistance, into the blebs instead of through the network. Local inhibition of blebbing is possible because opposite sides of the cells are effectively isolated from each other with respect to equilibration of pressure on a timescale of seconds.

Our poroelastic model could have important implications for other types of cell motility, because it implies that hydrostatic pressure can be generated and used locally to power shape change in animal cells. Leading-edge protrusion is typically coupled to actin polymerization[1–3], but hydrodynamic force could work together with polymerization force to power protrusion[7,8], and fluid flow could drive actin subunits forwards at a faster rate than that predicted by pure diffusion[28]. Localized hydrostatic pressure transients could be generated locally near the front of polarized cells by local recruitment and activation of myosin II as in blebbing, or of plasma membrane ion transporters (such as NHE1 (ref. 29)) that can trigger swelling through influx of an osmolyte (such as Na^+). Overall, our work shows that cytosolic fluid dynamics must be integrated with protein dynamics if we are to really understand cell motility.

METHODS

Local perfusion. Local perfusion was performed by injecting medium with the desired concentration of drug into a laminar fluid flow via a glass micropipette with a 1–2-μm diameter. Thin-walled borosilicate micropipettes (0.9-mm inner diameter) were pulled on a Sutter P87 pipette puller (Sutter instrument company). Coverslips on which cells had been cultured were affixed onto the bottom of the laminar fluid flow chamber (blueprint available upon request) with a 1:1:1 mix of vaseline, lanolin and paraffin. Laminar flow was created by feeding medium into the flow chamber at a constant rate of 30 ml h^{-1} with a syringe pump. A constant fluid level was maintained by aspirating the supernatant via a vacuum pump. The flow chamber consisted of a reservoir with a small opening (0.7 mm) that enables medium to enter the open part of the chamber and delivers a flow of approximately 30 μm s^{-1} in the vicinity of this entrance. For the flow to be sufficiently rapid to carry drug away, the target cells needed to be chosen within 200 μm of the reservoir opening. All experiments were performed at room temperature in Leibovitz-L15 medium with 10% 80:20 mix of donor calf serum:fetal calf serum. The microinjected solution was made up of medium with the desired concentration of drug and 0.1 μM of tetramethylrhodamine-isothiocyanate (TRITC, used as a fluorescent tracer) and filtered using a centrivac tube (Millipore). The micropipettes were backfilled with the solution and mounted onto a pipette holder attached to a three-axis oil hydraulic micromanipulator (Narishige). A pressure regulator (Eppendorf 5242, Eppendorf AG) set between 40 and 80 hPa was used to control flow from the micropipette.

Time-lapse video microscopy of the local perfusion experiments was performed as described in Supplementary Methods, except that in addition to DIC images, fluorescence images were also acquired (TRITC filters to visualize flow lines and fluorescein-isothiocyanate (FITC) filters to visualize WGA–Alexa

© 2005 Nature Publishing Group

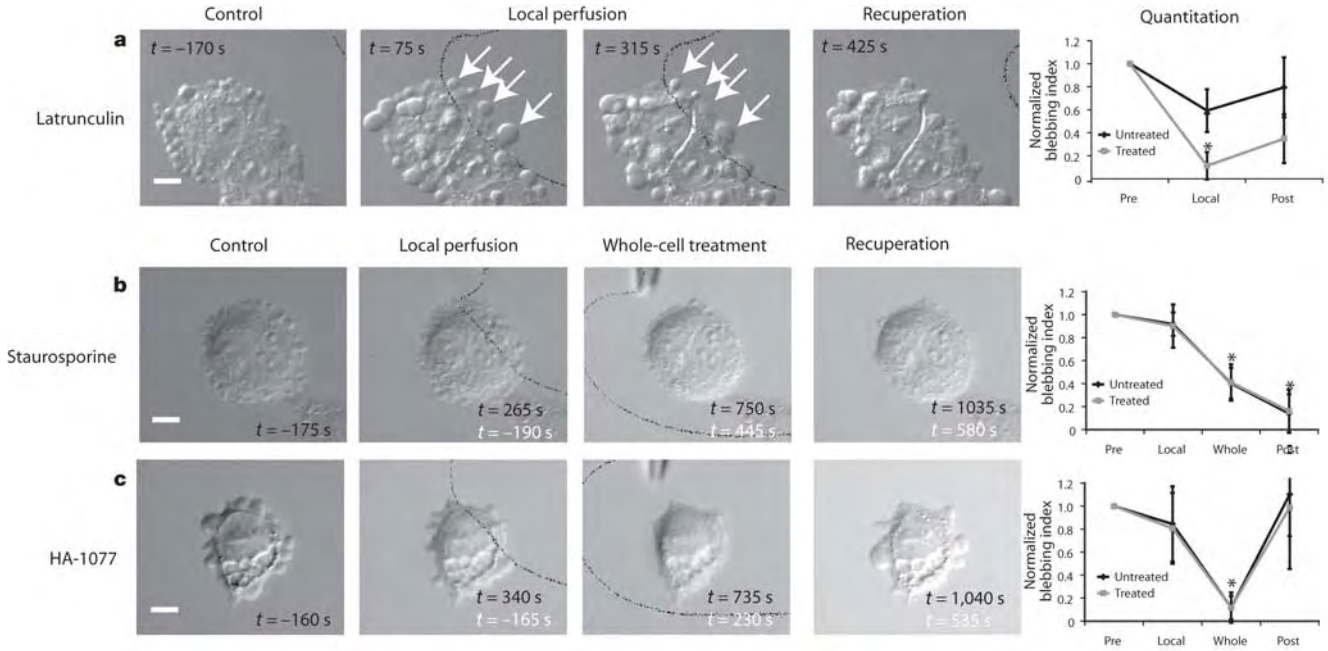


Figure 3 | Compounds with dual or global effects on blebbing. **a**, Local perfusion with latrunculin B has a dual effect. When latrunculin B is applied locally, bleb size increases globally ($t = 75$ s), and bleb dynamics cease in the treated area ($t = 75$ s and $t = 315$ s). Elsewhere, expansion and retraction continue. **b**, Global perfusion of staurosporine is necessary to inhibit blebbing. Before perfusion ($t = -175$ s) and after 395 s of local perfusion with staurosporine, blebbing is unperturbed throughout the cell. When staurosporine is applied to the whole cell (white text, $t = 445$ s), blebbing ceases. **c**, Global perfusion of HA-1077 is necessary to inhibit blebbing. After 340 s of local HA-1077 application, blebbing is unperturbed. When the whole cell is exposed to inhibitor ($t = 230$ s, white text), blebbing ceases. When perfusion is halted, blebbing is restored ($t = 535$ s, white text). Black lines delineate the flow coming out of the micropipette. Black text gives time relative to local application; white text gives time relative to global application. The normalized blebbing indices show the evolution over time of blebbing in the region exposed to inhibitor and in the free region. Error bars show the standard deviation. Asterisks denote significant changes in the blebbing index when compared with the initial blebbing index. Scale bars, 10 μm.

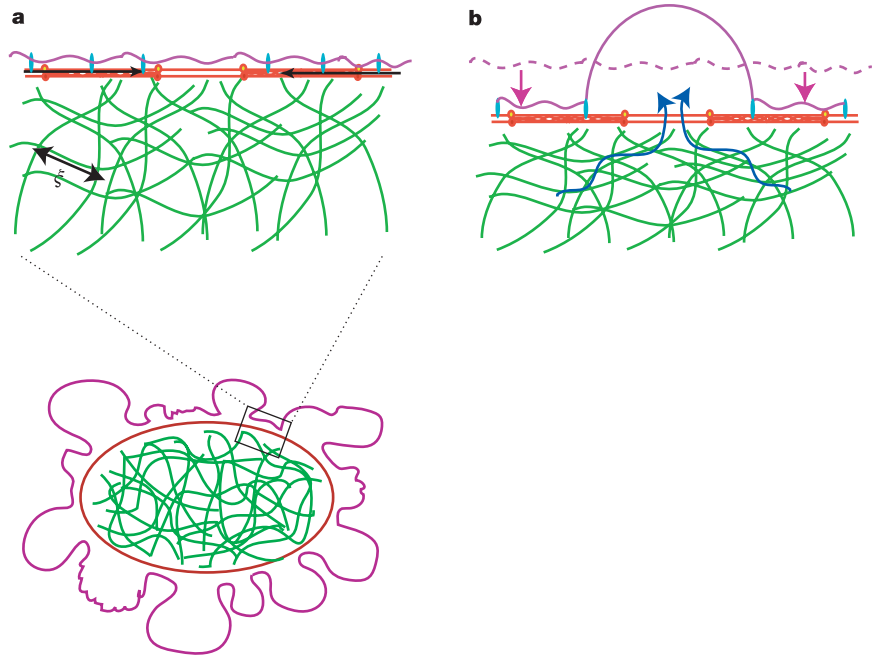


Figure 4 | Poroelastic description of blebbing. In all drawings, the actin cortex is drawn in red, the membrane is mauve and the cytoskeletal meshwork is green. **a**, In blebbing cells, a local contraction of myosin II (black arrows) associated with the actin cortex leads to a shortening of the cortical periphery and therefore to a compression of the cytoskeletal network that fills the cell. The cytoskeletal network is porous and has an average pore size ξ. The compression of the cytoskeletal meshwork creates a hydrostatic pressure in the vicinity of the region of contraction and can lead to bleb nucleation and expansion. **b**, Contraction of the actin cortex leads to a compression of the cytoskeletal network (the dashed mauve line indicates the original position of the cell surface) and drives flow of cytosol in the opposite direction (blue arrows). If there is a local defect in membrane–cytoskeleton attachment, a bleb is extruded. Bleb expansion is opposed by two forces: extracellular osmotic pressure and membrane tension.

© 2005 Nature Publishing Group

488). Exposure times were approximately 200 ms for DIC images, 600 ms for TRITC images and about 250 ms for FITC images. Images were acquired every 5 s with a × 20 objective using Metamorph (Universal Imaging Corp.). During a typical local perfusion experiment, about 30 images (approximately 150 s) were acquired before drug exposure. Then, the cell was exposed to drug for about 70 images (approximately 350 s), finally the cell was allowed to recover for 20 frames (about 100 s). When no local inhibition was observed, a whole-cell treatment was performed. This time, the cell was observed for 30 frames to ensure that it still blebbed, the whole cell was exposed to drug for 70 frames, and recovered for 20 frames. Initial concentrations of drug in the microinjected fluid were chosen to be two to three times those needed in bulk solution[25]. When no inhibition was observed when the whole cell was exposed to drug, the drug concentration in the micropipette was increased. When whole-cell inhibition was observed with only local perfusion, the drug concentration was decreased.

Post-hoc flow lines were superimposed onto the DIC images using Metamorph. The fluorescence images were assigned a threshold and intensities of 25% of the maximum intensity were converted to a binary image that was inverted and superimposed onto the DIC image using an AND operator.

Analysis and quantification of local perfusion assays. Each cell was divided into two regions: the treated region where local perfusion was applied, and the control untreated region. Each experiment was divided into different periods: (1) before the application of drug; (2) during local application of drug; (3) after application of drug; and, if need be, (4) during whole-cell treatment. The number of blebs occurring in each region for each period was counted manually. A blebbing index was computed for each region as follows: $B = \frac{N_{\text{blebs}}}{Lt}$, where N_{blebs} is the number of blebs observed during a given period t over a given perimeter L. These blebbing indices were used to compare the treated region to the untreated region before application of inhibitor, and the blebbing in either region during and after application of inhibitor to the blebbing in that region before application. Populations were compared with a Student's t-test and the level of significance was taken to be $P < 0.01$. For graphic output (Figs 2 and 3), the evolution of the blebbing indices was normalized to their values before application of local perfusion. In all cases, the blebbing index in the treated and untreated region was not significantly different before application of inhibitor ($P > 0.3$).

Received 19 January; accepted 15 March 2005.

1. Mahadevan, L. & Matsudaira, P. Motility powered by supramolecular springs and ratchets. *Science* **288**, 95–100 (2000).
2. Rafelski, S. M. & Theriot, J. A. Crawling toward a unified model of cell mobility: spatial and temporal regulation of actin dynamics. *Annu. Rev. Biochem.* **73**, 209–239 (2004).
3. Mogilner, A. & Oster, G. Polymer motors: pushing out the front and pulling up the back. *Curr. Biol.* **13**, R721–R733 (2003).
4. Messerli, M. A. & Robinson, K. R. Ionic and osmotic disruptions of the lily pollen tube oscillator: testing proposed models. *Planta* **217**, 147–157 (2003).
5. Money, N. P. & Harold, F. M. Extension growth of the water mold Achlya: interplay of turgor and wall strength. *Proc. Natl Acad. Sci. USA* **89**, 4245–4249 (1992).
6. Harold, F. M. Force and compliance: rethinking morphogenesis in walled cells. *Fungal Genet. Biol.* **37**, 271–282 (2002).
7. Tilney, L. G. & Inoue, S. Acrosomal reaction of the Thyone sperm. III. The relationship between actin assembly and water influx during the extension of the acrosomal process. *J. Cell Biol.* **100**, 1273–1283 (1985).
8. Condeelis, J. Life at the leading edge: the formation of cell protrusions. *Annu. Rev. Cell Biol.* **9**, 411–444 (1993).
9. Boal, D. H. *Mechanics of the Cell* (Cambridge Univ. Press, Cambridge, UK, 2002).
10. Drury, J. L. & Dembo, M. Hydrodynamics of micropipette aspiration. *Biophys. J.* **76**, 110–128 (1999).
11. Biot, M. General theory of three-dimensional consolidation. *J. Appl. Phys.* **12**, 155–164 (1941).
12. Wang, H. *Theory of Linear Poroelasticity with Applications to Geomechanics and Hydrogeology* (Princeton Univ. Press, Princeton, New Jersey, 2000).
13. Mills, J. C., Stone, N. L., Erhardt, J. & Pittman, R. N. Apoptotic membrane blebbing is regulated by myosin light chain phosphorylation. *J. Cell Biol.* **140**, 627–636 (1998).
14. Fishkind, D. J., Cao, L. G. & Wang, Y. L. Microinjection of the catalytic fragment of myosin light chain kinase into dividing cells: effects on mitosis and cytokinesis. *J. Cell Biol.* **114**, 967–975 (1991).
15. Burton, K. & Taylor, D. L. Traction forces of cytokinesis measured with optically modified elastic substrata. *Nature* **385**, 450–454 (1997).
16. Trinkaus, J. P. Surface activity and locomotion of Fundulus deep cells during blastula and gastrula stages. *Dev. Biol.* **30**, 69–103 (1973).
17. Friedl, P. & Wolf, K. Tumour-cell invasion and migration: diversity and escape mechanisms. *Nature Rev. Cancer* **3**, 362–374 (2003).
18. Albrecht-Buehler, G. Does blebbing reveal the convulsive flow of liquid and solutes through the cytoplasmic meshwork? *Cold Spring Harb. Symp. Quant. Biol.* **46**, 45–49 (1982).
19. Cunningham, C. C. Actin polymerization and intracellular solvent flow in cell surface blebbing. *J. Cell Biol.* **129**, 1589–1599 (1995).
20. Cunningham, C. C. *et al.* Actin-binding protein requirement for cortical stability and efficient locomotion. *Science* **255**, 325–327 (1992).
21. Cheung, A. *et al.* A small-molecule inhibitor of skeletal muscle myosin II. *Nature Cell Biol.* **4**, 83–88 (2002).
22. Dai, J. & Sheetz, M. P. Membrane tether formation from blebbing cells. *Biophys. J.* **77**, 3363–3370 (1999).
23. Evans, E. & Leung, A. Adhesivity and rigidity of erythrocyte membrane in relation to wheat germ agglutinin binding. *J. Cell Biol.* **98**, 1201–1208 (1984).
24. Popov, S., Brown, A. & Poo, M. M. Forward plasma membrane flow in growing nerve processes. *Science* **259**, 244–246 (1993).
25. O'Connell, C. B., Warner, A. K. & Wang, Y. Distinct roles of the equatorial and polar cortices in the cleavage of adherent cells. *Curr. Biol.* **11**, 702–707 (2001).
26. Yarrow, J. C., Totsukawa, G., Charras, G. T. & Mitchison, T. J. Screening for cell migration inhibitors via automated microscopy reveals a rho-kinase inhibitor. *Chem. Biol.* **12**, 385–395 (2005).
27. Takayama, S. *et al.* Selective chemical treatment of cellular microdomains using multiple laminar streams. *Chem. Biol.* **10**, 123–130 (2003).
28. Zicha, D. *et al.* Rapid actin transport during cell protrusion. *Science* **300**, 142–145 (2003).
29. Baumgartner, M., Patel, H. & Barber, D. L. Na(+)/H(+) exchanger NHE1 as plasma membrane scaffold in the assembly of signaling complexes. *Am. J. Physiol. Cell Physiol.* **287**, C844–C850 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors would like to acknowledge the Nikon Imaging Centre at Harvard Medical School and, in particular, J. Waters. The authors would also like to acknowledge J. Horn at the HMS machine shop for manufacturing the perfusion chamber. G.T.C. was in receipt of a Wellcome Trust Overseas Fellowship. M.A.H. is supported by a Programme Grant from the Wellcome Trust. L.M. was supported by NSF-MRSEC at Harvard University. This work was supported by a grant from the NIH to T.J.M.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.T.C. (gcharras@hms.harvard.edu).

© 2005 Nature Publishing Group

LETTERS

DNA synthesis provides the driving force to accelerate DNA unwinding by a helicase

Natalie M. Stano[1], Yong-Joo Jeong[1,2], Ilker Donmez[1], Padmaja Tummalapalli[1], Mikhail K. Levin[3] & Smita S. Patel[1]

Helicases are molecular motors that use the energy of nucleoside 5′-triphosphate (NTP) hydrolysis to translocate along a nucleic acid strand and catalyse reactions such as DNA unwinding. The ring-shaped helicase[1] of bacteriophage T7 translocates along single-stranded (ss)DNA at a speed of 130 bases per second[2]; however, T7 helicase slows down nearly tenfold when unwinding the strands of duplex DNA[3]. Here, we report that T7 DNA polymerase, which is unable to catalyse strand displacement DNA synthesis by itself, can increase the unwinding rate to 114 base pairs per second, bringing the helicase up to similar speeds compared to its translocation along ssDNA. The helicase rate of stimulation depends upon the DNA synthesis rate and does not rely on specific interactions between T7 DNA polymerase and the carboxy-terminal residues of T7 helicase. Efficient duplex DNA synthesis is achieved only by the combined action of the helicase and polymerase. The strand displacement DNA synthesis by the DNA polymerase depends on the unwinding activity of the helicase, which provides ssDNA template. The rapid trapping of the ssDNA bases by the DNA synthesis activity of the polymerase in turn drives the helicase to move forward through duplex DNA at speeds similar to those observed along ssDNA.

The DNA factory of bacteriophage T7 is one of the simplest and most widely used model systems for studying replication mechanisms[4]. The T7 replication complex, containing a helicase (T7 gp4), a DNA polymerase (T7 gp5 in complex with *Escherichia coli* thioredoxin) and a ssDNA-binding protein (T7 gp2.5), efficiently catalyses leading and lagging strand DNA synthesis[5]. T7 DNA polymerase (polymerase-thioredoxin) alone can elongate a DNA primer when the downstream DNA template is single stranded (Fig. 1a). The average rate of DNA synthesis by T7 DNA polymerase increases in a hyperbolic manner with dNTP concentration, with a $K_{1/2}$ of 11 μM and V_{max} of 230 nucleotides per second at 18 °C (Fig. 1b), which is consistent with previous pre-steady-state kinetic measurements[6]. DNA synthesis is blocked when the downstream template DNA is duplex (Fig. 1c). T7 DNA polymerase incorporates only 4–5 nucleotides on the duplex template before DNA synthesis stalls. These results indicate that T7 DNA polymerase cannot unwind the duplex DNA beyond 4–5 base pairs (bp) and hence cannot catalyse strand displacement DNA synthesis.

T7 helicase uses the energy of dTTP hydrolysis for translocation and unwinding of duplex DNA[3,7–9]. Using an all-or-none radiometric assay carried out under single-turnover conditions[10], we measured the unwinding activity of T7 helicase on a 30-bp replication substrate (Fig. 2a, b). T7 helicase was pre-incubated with replication substrate (Fig. 2a) in the presence of dTTP without Mg^{2+} (conditions that allow assembly of the protein on the DNA, but no unwinding), and the reaction was started by rapid addition of Mg^{2+}. T7 helicase unwinds the replication substrate at an average rate of 9 bp s^{-1} in the absence of T7 DNA polymerase (Fig. 2b). In the presence of T7 DNA polymerase, saturating dTTP and with the other three dNTPs at 100 μM—conditions under which T7 DNA polymerase is capable of incorporating nucleotides at its maximal rate—the unwinding rate of T7 helicase was 114 bp s^{-1} (Fig. 2b, c). This is about a 13-fold enhancement in the unwinding rate of T7 helicase.

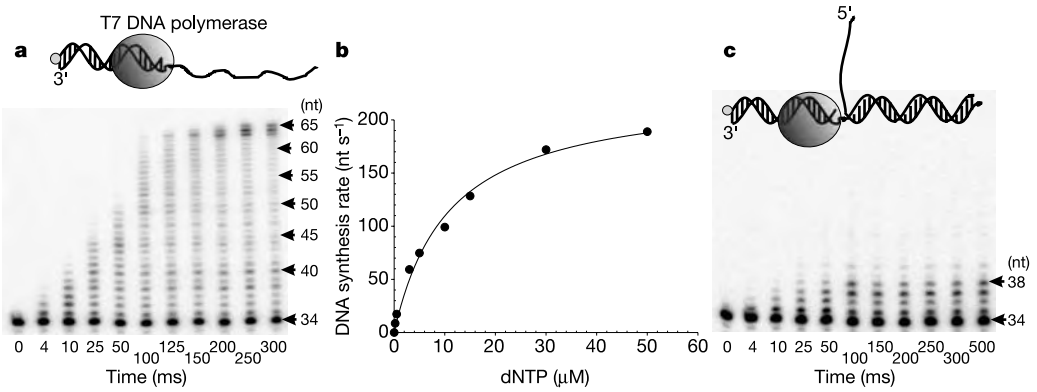


Figure 1 | DNA synthesis by T7 DNA polymerase. a, Progressive elongation of the radiolabelled 34-nucleotide primer annealed to the 65-nucleotide template at a dNTP concentration of 100 μM. **b**, The average DNA synthesis rate increases hyperbolically with dNTP concentration (observed rate $= V_{max}[dNTP]/(K_{1/2} + [dNTP])$ with $K_{1/2} = 11 \pm 2$ μM and $V_{max} = 228 \pm 11$ nucleotides per second. **c**, Elongation of the 34-nucleotide primer on the 30-bp replication substrate. nt, nucleotides.

[1]Department of Biochemistry, UMDNJ-Robert Wood Johnson Medical School, 675 Hoes Lane, Piscataway, New Jersey 08854, USA. [2]Department of Bio and Nanochemistry, Kookmin University, 861-1, Chongnung-dong, Songbuk-gu, Seoul 136-702, Korea. [3]Center for Cell Analysis and Modeling, University of Connecticut Health Center, Farmington, Connecticut 06030-1507, USA.

© 2005 Nature Publishing Group

We find that the helicase rate of stimulation depends on the rate of DNA synthesis, which in turn depends on the concentration of dNTP (Fig. 1b). T7 helicase uses only dTTP as an energy source for unwinding (dissociation constant (K_d) of approximately 80 μM)[3], and other dNTPs do not support DNA unwinding[8]. To investigate the effect of DNA synthesis rate on helicase rate of stimulation, we varied the concentration of three dNTPs while keeping dTTP at 1 mM. The observed unwinding rate increases with increasing dNTP concentration in a hyperbolic fashion, with a $K_{1/2}$ identical to that for DNA synthesis (Fig. 1b) and V_{max} of 130 bp s^{-1} (Fig. 2c). These results show that the unwinding rate of T7 helicase depends on the speed of T7 DNA polymerase, and when the DNA polymerase can incorporate nucleotides at a rate much faster than the translocation rate of the helicase, it stimulates the helicase to unwind DNA at the same rate as its translocation along ssDNA (130 nt s^{-1})[2].

The coupled action of the helicase and polymerase can be monitored by the primer elongation assay. We show that the 30-bp DNA substrate is not replicated efficiently by T7 DNA polymerase alone (Fig. 1c); however, T7 DNA polymerase replicates this substrate efficiently in the presence of T7 helicase (Fig. 3a). The time course of base additions shows that about 18 nucleotides are added in 150 ms. This corresponds to a maximum rate of DNA synthesis of about 120 nucleotides per second by the helicase–polymerase complex. This rate is comparable to the one measured by the all-or-none unwinding assay (Fig. 2b, c). Thus, both the DNA unwinding assay and the DNA synthesis assay show that T7 DNA polymerase activates T7 helicase.

The activities of T7 helicase and T7 DNA polymerase are interdependent. The helicase provides the DNA polymerase with the ssDNA template for DNA synthesis. Similarly, the DNA polymerase serves to increase the unwinding rate of the helicase. In addition, the experiments show that the rate of DNA synthesis dictates the stimulated unwinding rate. What is the mechanism of this stimulation? Because T7 helicase and T7 DNA polymerase specifically interact to form a complex, one possibility is that the rate increase occurs through an allosteric change in the helicase producing a better unwinding motor.

T7 helicase forms a specific complex with T7 DNA polymerase through 17 carboxy-terminal amino acid residues[11]. To test the function of the specific complex formation, we prepared a mutant T7 helicase that lacked these 17 C-terminal residues (ΔCt mutant). The ΔCt mutant of T7 helicase has both wild-type ssDNA-stimulated dTTPase and helicase activities[11], translocates along ssDNA with a rate similar to wild-type helicase (data not shown), but does not form a stable complex with the polymerase[11]. Rapid DNA synthesis was observed in the presence of the ΔCt helicase mutant (Fig. 3b). On the basis of the addition of 16 nucleotides in 100 ms, we estimate that the 30-bp duplex is replicated with a rate of about 160 nucleotides per

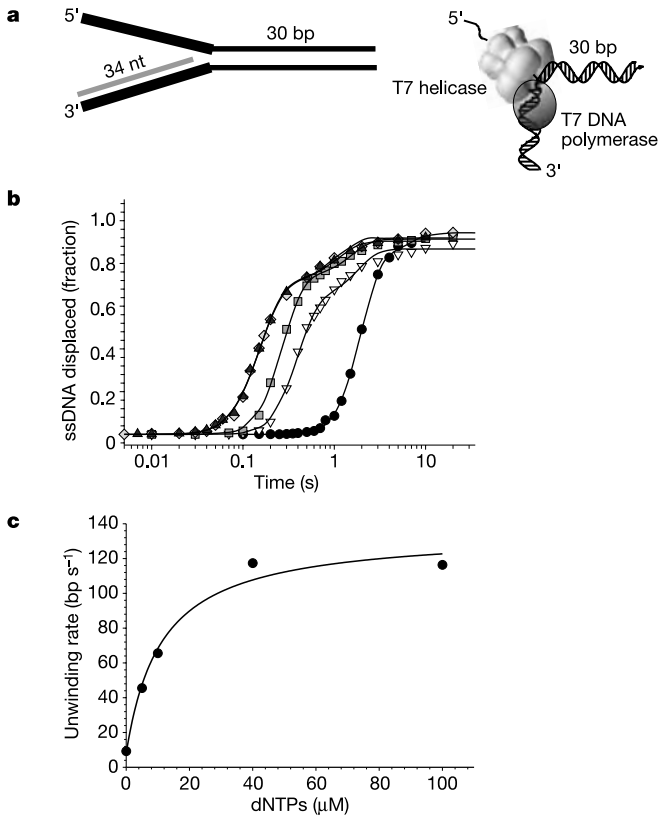


Figure 2 | The unwinding rate of T7 helicase depends on the speed of DNA synthesis. **a**, 30-bp replication substrate. **b**, Unwinding kinetics of the 30-bp region of the replication substrate in the absence (circles) or presence of T7 DNA polymerase at 1 mM dTTP and at 5 μM (inverted white triangles), 10 μM (squares), 40 μM (filled triangles) and 100 μM (diamonds) for each of the other three dNTPs. An unwinding rate (sk) of 9.3 bp s^{-1} (80% amplitude) and 2.6 bp s^{-1} (20%) was obtained in the absence of T7 DNA polymerase by fitting the kinetics to equation (1) ($s = 3.6 \pm 0.15$ bp $step^{-1}$; $k_1 = 2.6 \pm 0.1$ steps s^{-1}; $k_2 = 0.7 \pm 0.1$ steps s^{-1}). In the presence of T7 DNA polymerase and a dNTP concentration of 100 μM, the unwinding rate is 114 bp s^{-1} (90%) and 15 bp s^{-1} (10%) ($s = 5.3 \pm 0.4$ bp $step^{-1}$; $k_1 = 22 \pm 1.6$ steps s^{-1}; $k_2 = 3 \pm 0.4$ steps s^{-1}). **c**, The unwinding rate increases hyperbolically with dNTP concentration, with $K_{1/2} = 11 \pm 4$ μM, and provides a maximum unwinding rate of 127 ± 12 bp s^{-1}.

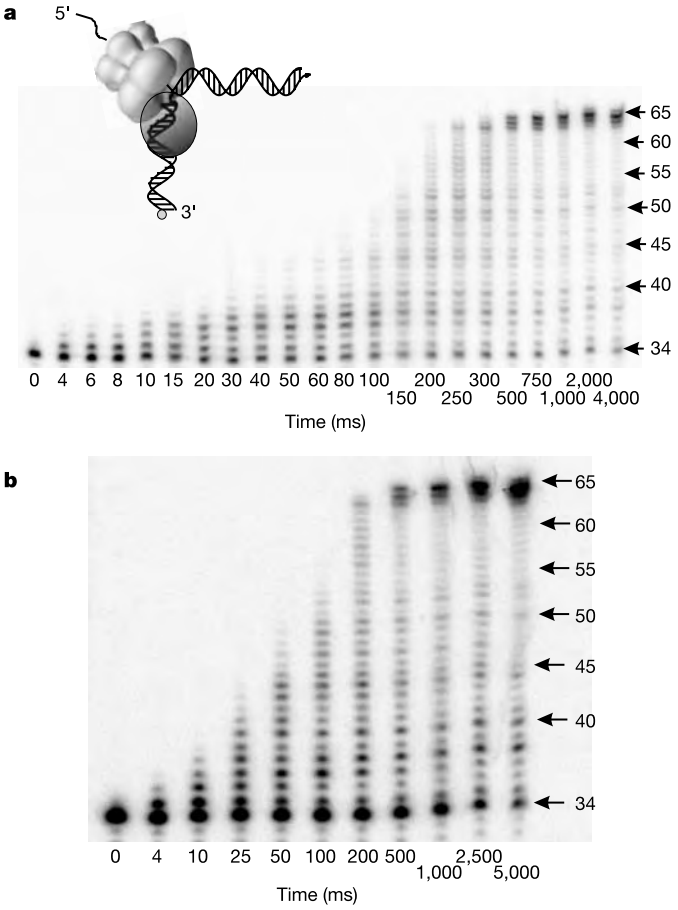


Figure 3 | DNA synthesis by T7 DNA polymerase and T7 helicase. **a**, Progressive elongation of the 34-nucleotide primer by T7 DNA polymerase and T7 helicase on the 30-bp replication substrate. **b**, Kinetics of DNA synthesis by T7 DNA polymerase and the ΔCt T7 helicase mutant on the 30-bp replication substrate.

© 2005 Nature Publishing Group

second. Therefore, the interaction between T7 DNA polymerase and the C-terminal residues of T7 helicase is not necessary for stimulation of the helicase reaction. Of note, the interaction between T7 helicase and T7 DNA polymerase serves to increase the processivity of the complex, which becomes critical for copying long DNA templates[11]. T7 helicase and T7 DNA polymerase possess a sufficient degree of intrinsic processivity to unwind and replicate the short DNA duplexes used here, in the absence of specific interactions involving the C-terminal residues of the helicase. A smaller fraction of DNA primer was elongated initially with the ΔCt T7 helicase, indicating that specific interactions also have a function in the initial complex assembly.

Any protein that binds tightly to ssDNA can shift the equilibrium of the DNA unwinding reaction (dsDNA ⇌ two ssDNAs) towards ssDNA[12]. Therefore, the DNA polymerase might facilitate the helicase unwinding activity by trapping the complementary strand that is displaced by the helicase. Consequently, we investigated whether ssDNA-binding proteins would have the same effect. We used a fork DNA as an unwinding substrate that contained a T_{15} 3′ ssDNA tail, as well as T7 gp2.5 (ref. 13) and *E. coli* SSB[14] proteins that bind ssDNA with high affinity. The results (Supplementary Fig. 1) show that the ssDNA-binding proteins have little effect on the unwinding rate. A modest increase in the initial unwinding amplitude was observed, which suggests that the processivity of unwinding is higher in the presence of the ssDNA-binding proteins.

The experiments with ssDNA-binding proteins indicate that helicase stimulation by T7 DNA polymerase must involve more than simply binding of the protein to the displaced complementary strand. Although both DNA polymerase and SSB proteins capture the nascent DNA strand, the polymerase does so more rapidly and with an increment of one base. The fact that stimulation of T7 helicase depends on the speed of the DNA polymerase substantiates this hypothesis. Our model predicts that a heterologous DNA polymerase that is as fast and processive as T7 DNA polymerase should also be able to stimulate the unwinding rate of T7 helicase. Indeed, this was observed in a study where T7 DNA polymerase was shown to support efficient duplex DNA replication with the bacteriophage T4 helicase[15].

To test the ability of a heterologous polymerase to stimulate unwinding, we used T4 gp43 core DNA polymerase, which does not have strand displacement activity (Fig. 4a). In the presence of T7 helicase, T4 DNA polymerase catalyses strand displacement DNA synthesis at a rate of about 30 nucleotides per second (Fig. 4b). Thus, T4 DNA polymerase provides a threefold rate increase. The lower degree of stimulation is consistent with the slower speed of T4 DNA polymerase on a ssDNA template (about 50 nucleotides per second) (Fig. 4c). To compensate for the different intrinsic synthesis rates, the optimal stimulated rate of the T4 polymerase–helicase should be compared with the rate of T7 polymerase–helicase measured at 1.8 μM (approximately 50–60 nucleotides per second, compare Fig. 4e and c), the conditions providing the same intrinsic synthesis rate. Under these conditions, T7 DNA polymerase stimulates T7 helicase to the same extent as T4 DNA polymerase (compare Fig. 4d and b). The fraction of DNA primer elongated to full-length product is less with the heterologous T4 DNA polymerase, consistent with the idea that specific interactions are important for initial complex assembly. These results indicate that a heterologous DNA polymerase and helicase can show cooperativity in DNA synthesis and unwinding, and that the extent of the helicase activity stimulation depends upon the intrinsic rate of DNA synthesis.

The reduced translocation rate of the helicase during unwinding and its stimulation by polymerases can be explained in terms of a brownian motor mechanism[16–18]. The brownian motor (helicase) translocates along the DNA strand through a succession of power strokes and diffusion states. While in the diffusion state the helicase can move in either direction along the DNA; the power strokes make the net movement unidirectional. In the presence of the complimentary strand, DNA base pairs fluctuating between closed and open states produce a net force directed against the helicase forward motion. The power strokes of the helicase can overcome the force; however, during the diffusion phase, backward motion occurs frequently. Therefore, the observed DNA unwinding rate is slower than the helicase's intrinsic translocation rate along ssDNA[3].

T7 and T4 DNA polymerases lack strand displacement activity (unwinding) beyond 4–5 bp. Thus, the activities of these DNA polymerases are dependent on the unwinding activity of the helicase, which provides a ssDNA template. The action of the DNA polymerase (that is, the formation of new duplex DNA) in turn results in efficient trapping of the DNA strand created by the helicase. The forward steps of the DNA polymerase may serve to increase the helicase's net forward movement (push) or decrease the helicase's backward slips (brake). The observed unwinding rate depends on the speed of DNA synthesis. Only when the polymerase rapidly traps the ssDNA bases as soon as they are formed by the helicase can the helicase obtain its maximal rate, and the complex is able to move through duplex DNA at speeds similar to those observed for the helicase along ssDNA.

The synergy between helicase and polymerase is important in DNA replication, and is also evident from studies of replication complexes of *E. coli*[19,20], human mitochondria[21] and bacteriophage T4[15,22,23]. Studies of the simpler T7 system continue to provide a deeper understanding of the enzymatic mechanisms of DNA replication, many of which are likely to be generally applicable to the more complex replication systems.

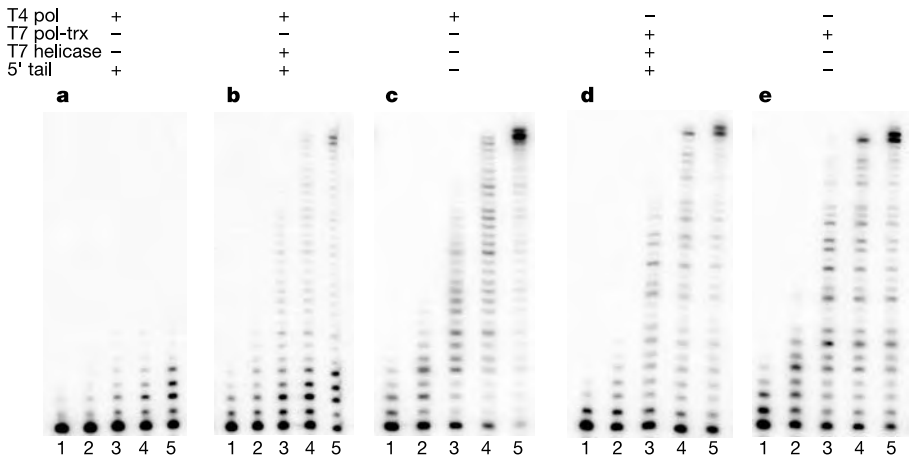


Figure 4 | DNA synthesis by T4 and T7 DNA polymerases and T7 helicase. Lanes 1–5 in all experiments correspond to reaction times of 0.025, 0.1, 0.5, 1.0 and 3.0 s, respectively. **a**, DNA synthesis by T4 DNA polymerase on the 30-bp replication substrate at 100 μM for each dNTP. **b**, DNA synthesis by T4 DNA polymerase and T7 helicase on the replication substrate, at 1 mM dTTP and 100 μM for each of the other three dNTPs. **c**, DNA synthesis by T4 DNA polymerase on the substrate with a ssDNA template at 100 μM for all dNTPs. **d**, DNA synthesis by T7 DNA polymerase (polymerase-thioredoxin or pol-trx) and T7 helicase on the replication substrate at 1 mM dTTP and 1.8 μM for each of the three other dNTPs. **e**, DNA synthesis by T7 DNA polymerase on the substrate with a ssDNA template at 1.8 μM for each dNTP.

© 2005 Nature Publishing Group

METHODS

DNA. Oligodeoxynucleotides (Integrated DNA Technologies) were purified by polyacrylamide gel electrophoresis (PAGE), and DNA concentrations were determined in 8 M urea (260 nm absorbance and calculated extinction coefficients). DNA was radiolabelled at the 5′ end using [γ-^{32}P]ATP and T4 polynucleotide kinase. The replication substrate was made by annealing the 5′ strand (5′-T_{35}-GAGCGGATTACTATACTACATTAGAATTCA), 34-nucleotide primer (5′-CTAGTTACAGAGTTATGGTGACGATACAAACTAT) and the 3′ strand (3′-GATCAATGTCTCAATACCACTGCTATGTTTGATATCTCGCC TAATGATATGATGTAATCTTAAGT). The fork DNA was made by annealing the 5′ strand with the T_{15} 3′ strand (3′-T_{15}-CTCGCCTAATGATATGATGT AATCTTAAGT).

Proteins. T7 helicase gp4A′, a M64L mutant of T7 helicase–primase protein, was purified as described previously[9]. T7 gp5 (D5A, E7A, exo$^-$ gp5) was purified as described[6] and *E. coli* thioredoxin was purchased from Sigma Chemicals. The ΔCt T7 helicase was prepared as described[11]. T7 gp2.5 was purified as described[13], *E. coli* SSB was a gift from M. O'Donnell, and T4 DNA polymerase exo$^-$ mutant (D219A)[24] was a gift from A. Berdis. Protein concentrations were determined by absorbance measurements at 280 nm in 8 M urea using their calculated extinction coefficients.

DNA unwinding kinetics. Unwinding kinetics were measured in a RQF3 rapid quench-flow instrument (KinTek) at 18 °C (ref. 3). T7 helicase, DNA substrate (with radiolabelled 5′-strand), 2 mM dTTP and 3 mM EDTA in buffer T (50 mM Tris-Cl, pH 7.5, 40 mM NaCl, 10% glycerol) was loaded in one syringe of the quench-flow apparatus. $MgCl_2$ (final free concentration of 4 mM), dNTPs (various concentrations) and trap (3 μM of unlabelled 5′ strand) was added from the second syringe to initiate the reaction. The reactions were quenched after pre-determined times, resolved by native PAGE, and products analysed as described previously[3]. The unwinding reactions contained a final 200 nM T7 helicase hexamer and 2.5 nM of 30-bp replication substrate. The unwinding reactions with 200 nM T7 DNA polymerase contained 2 μM *E. coli* thioredoxin, 200 nM T7 helicase and 100 nM replication substrate. Unwinding reactions with ssDNA-binding protein contained 5 μM T7 gp2.5 or 1 μM of *E. coli* SSB, and these were added with $MgCl_2$ at the start of the reaction. No DNA trap was added with the $MgCl_2$ when ssDNA-binding protein was present, but the trap was added with the quenching solution to prevent the unwound strands from re-annealing.

Data analysis. Kinetic fitting was performed using MATLAB with Optimization toolbox software (MathWorks). Stepwise unwinding kinetics were described using the incomplete gamma function[3,10,25]. Best fits were obtained assuming unwinding by two populations of helicase species with identical step size *s*, but different stepping rates (k_1 and k_2). The stepping kinetics are described by equation (1):

$$F = A_1\Gamma(k_1, t, n) + A_2\Gamma(k_2, t, n) \quad (1)$$

where F is the fraction of unwound DNA substrate molecules, A_1 and A_2 are the amplitudes of unwinding, the incomplete gamma function

$$\Gamma(k, t, n) = \frac{1}{\int_0^\infty e^{-x}x^{n-1}\mathrm{d}x}\int_0^{kt} e^{-x}x^{n-1}\mathrm{d}x,$$

and t is reaction time. The number of steps $n = (L - L_m)/s$, where s is step size, L is the number of base pairs in the DNA substrate duplex and L_m is the length of the shortest DNA duplex that can stay together under the experimental conditions, which was fixed at 12 bp, based on previous studies[3].

DNA synthesis kinetics. DNA synthesis kinetics was measured using the rapid quench-flow instrument at 18 °C. The final concentrations are reported in parentheses. The DNA substrate with a radiolabelled 34-nucleotide primer (100 nM) was incubated with the DNA polymerase (200 nM) and *E. coli* thioredoxin (2 μM with T7 DNA polymerase) with or without T7 helicase or ΔCt T7 helicase (200 nM hexamer) in buffer T containing dTTP (1 mM) and EDTA (1.5 mM). This solution was mixed rapidly with dNTPs (various concentrations) and $MgCl_2$ (free final Mg^{2+} concentration was 4 mM) in buffer T to initiate the reaction. After pre-determined times, reactions were quenched (200 mM EDTA), resolved by electrophoresis on a 14% or 16% polyacrylamide/7 M urea gel (0.25–0.75-mm wedge spacers), and visualized and quantified using the PhosphorImager (ImageQuant software). The average rate of DNA synthesis was determined from the exponential increase in DNA products with time, $D_P = A(1 - e^{-kt})$, where D_P is the concentration of DNA products, A is the amplitude, t is time and k is the observed DNA synthesis rate.

Received 28 August 2004; accepted 29 March 2005.

1. Patel, S. S. & Picha, K. M. Structure and function of hexameric helicases. *Annu. Rev. Biochem.* **69,** 651–697 (2000).
2. Kim, D. E., Narayan, M. & Patel, S. S. T7 DNA helicase: a molecular motor that processively and unidirectionally translocates along single-stranded DNA. *J. Mol. Biol.* **321,** 807–819 (2002).
3. Jeong, Y. J., Levin, M. K. & Patel, S. S. The DNA-unwinding mechanism of the ring helicase of bacteriophage T7. *Proc. Natl Acad. Sci. USA* **101,** 7264–7269 (2004).
4. Richardson, C. C. Bacteriophage T7: minimal requirements for the replication of a duplex DNA molecule. *Cell* **33,** 315–317 (1983).
5. Lee, J., Chastain, P. D., Kusakabe, T., Griffith, J. D. & Richardson, C. C. Coordinated leading and lagging strand DNA synthesis on a minicircular template. *Mol. Cell* **1,** 1001–1010 (1998).
6. Patel, S. S., Wong, I. & Johnson, K. A. Pre-steady-state kinetic analysis of processive DNA replication including complete characterization of an exonuclease-deficient mutant. *Biochemistry* **30,** 511–525 (1991).
7. Matson, S. W. & Richardson, C. C. DNA-dependent nucleoside 5′-triphosphatase activity of the gene 4 protein of bacteriophage T7. *J. Biol. Chem.* **258,** 14009–14016 (1983).
8. Matson, S. W., Tabor, S. & Richardson, C. C. The gene 4 protein of bacteriophage T7. Characterization of helicase activity. *J. Biol. Chem.* **258,** 14017–14024 (1983).
9. Patel, S. S., Rosenberg, A. H., Studier, F. W. & Johnson, K. A. Large scale purification and biochemical characterization of T7 primase/helicase proteins. Evidence for homodimer and heterodimer formation. *J. Biol. Chem.* **267,** 15013–15021 (1992).
10. Ali, J. A. & Lohman, T. M. Kinetic measurement of the step size of DNA unwinding by *Escherichia coli* UvrD helicase. *Science* **275,** 377–380 (1997).
11. Notarnicola, S. M., Mulcahy, H. L., Lee, J. & Richardson, C. C. The acidic carboxyl terminus of the bacteriophage T7 gene 4 helicase/primase interacts with T7 DNA polymerase. *J. Biol. Chem.* **272,** 18425–18433 (1997).
12. von Hippel, P. H. & Delagoutte, E. A general model for nucleic acid helicases and their "coupling" within macromolecular machines. *Cell* **104,** 177–190 (2001).
13. Kim, Y. T., Tabor, S., Bortner, C., Griffith, J. D. & Richardson, C. C. Purification and characterization of the bacteriophage T7 gene 2.5 protein. A single-stranded DNA-binding protein. *J. Biol. Chem.* **267,** 15022–15031 (1992).
14. Lohman, T. M. & Ferrari, M. E. *Escherichia coli* single-stranded DNA-binding protein: multiple DNA-binding modes and cooperativities. *Annu. Rev. Biochem.* **63,** 527–570 (1994).
15. Delagoutte, E. & von Hippel, P. H. Molecular mechanisms of the functional coupling of the helicase (gp41) and polymerase (gp43) of bacteriophage T4 within the DNA replication fork. *Biochemistry* **40,** 4459–4477 (2001).
16. Levin, M. K. & Patel, S. S. in *Molecular Motors* (ed. Schliwa, M.) 179–198 (Wiley, Weinheim, 2003).
17. Levin, M. K., Gurjar, M. M. & Patel, S. S. ATP binding modulates the nucleic acid affinity of hepatitis C virus helicase. *J. Biol. Chem.* **278,** 23311–23316 (2003).
18. Wang, H. & Oster, G. Ratchets, power strokes, and molecular motors. *Applied Phys. A* **75,** 315–323 (2002).
19. Kim, S., Dallmann, H. G., McHenry, C. S. & Marians, K. J. Coupling of a replicative polymerase and helicase: a tau-DnaB interaction mediates rapid replication fork movement. *Cell* **84,** 643–650 (1996).
20. Mok, M. & Marians, K. J. The *Escherichia coli* preprimosome and DNA B helicase can form replication forks that move at the same rate. *J. Biol. Chem.* **262,** 16644–16654 (1987).
21. Korhonen, J. A., Pham, X. H., Pellegrini, M. & Falkenberg, M. Reconstitution of a minimal mtDNA replisome *in vitro*. *EMBO J.* **23,** 2423–2429 (2004).
22. Schrock, R. D. & Alberts, B. Processivity of the gene 41 DNA helicase at the bacteriophage T4 DNA replication fork. *J. Biol. Chem.* **271,** 16678–16682 (1996).
23. Cha, T. A. & Alberts, B. M. The bacteriophage T4 DNA replication fork. Only DNA helicase is required for leading strand DNA synthesis by the DNA polymerase holoenzyme. *J. Biol. Chem.* **264,** 12220–12225 (1989).
24. Frey, M. W., Nossal, N. G., Capson, T. L. & Benkovic, S. J. Construction and characterization of a bacteriophage T4 DNA polymerase deficient in 3′ → 5′ exonuclease activity. *Proc. Natl Acad. Sci. USA* **90,** 2579–2583 (1993).
25. Lucius, A. L., Maluf, N. K., Fischer, C. J. & Lohman, T. M. General methods for analysis of sequential "n-step" kinetic mechanisms: Application to single turnover kinetics of helicase-catalyzed DNA unwinding. *Biophys. J.* **85,** 2224–2239 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. O'Donnell, A. Berdis, N. Andraos, C. C. Richardson and M. Salas for the gift of proteins, and C. M. Drain for critical reading of the manuscript. This research was supported by an NIH grant to S.S.P.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.S.P (patelss@umdnj.edu).

© 2005 Nature Publishing Group

LETTERS

Structure of the zinc-binding domain of an essential component of the hepatitis C virus replicase

Timothy L. Tellinghuisen[1], Joseph Marcotrigiano[1] & Charles M. Rice[1]

Hepatitis C virus (HCV) is a human pathogen affecting nearly 3% of the world's population[1]. Chronic infections can lead to cirrhosis and liver cancer. The RNA replication machine of HCV is a multi-subunit membrane-associated complex. The non-structural protein NS5A is an active component of HCV replicase[2,3], as well as a pivotal regulator of replication[2,4] and a modulator of cellular processes ranging from innate immunity to dysregulated cell growth[5,6]. NS5A is a large phosphoprotein (56–58 kDa) with an amphipathic α-helix at its amino terminus that promotes membrane association[7–9]. After this helix region, NS5A is organized into three domains[10]. The N-terminal domain (domain I) coordinates a single zinc atom per protein molecule[10]. Mutations disrupting either the membrane anchor[7,8] or zinc binding[10] of NS5A are lethal for RNA replication. However, probing the role of NS5A in replication has been hampered by a lack of structural information about this multifunctional protein. Here we report the structure of NS5A domain I at 2.5-Å resolution, which contains a novel fold, a new zinc-coordination motif and a disulphide bond. We use molecular surface analysis to suggest the location of protein-, RNA- and membrane-interaction sites.

The structure of NS5A domain I (amino acids 36–198) reveals two identical monomers per asymmetric unit, packed as a dimer via contacts near the N-terminal ends of the molecules. For ease of discussion, we divide domain I into two subdomains: the N-terminal subdomain IA and the C-terminal subdomain IB (Fig. 1). A more schematic view of the fold is shown in Fig. 1e. Using the DALI[11] server, no structures were found related to domain I or either subdomain, indicating that this protein represents a novel fold. Subdomain IA consists of an N-terminal extended loop lying adjacent to a three-stranded anti-parallel β-sheet (strands B1–B3), with α-helix H2 (designated H2 to allow numbering of the N-terminal membrane-anchoring helix as H1, ref. 8) at the C terminus of the third β-strand. These elements comprise the structural scaffold for a four-cysteine zinc-atom-coordination site at one end of the β-sheet.

Figures 2a and b show a view of subdomain IA (amino acids 36–100), highlighting the cysteine residues involved in zinc coordination. The anti-parallel β-sheet (strands B1–B3), which positions Cys 57, Cys 59 and Cys 80 near the zinc-binding site, is essentially as predicted in our previous model[10]. The long random coil region, which positions Cys 39 and connects the N terminus to the β-sheet had been incorrectly predicted as a β-strand, but the arrangement of this region in relation to the other strands matches the original model. The distances observed between the cysteine side-chain sulphur groups and the zinc atom for Cys 39 (2.36 Å), Cys 57 (2.47 Å), Cys 59 (2.42 Å) and Cys 80 (2.45 Å) are close to the ideal distance (2.35 ± 0.09 Å) for structural metal zinc-coordination sites[12]. Similarly, the side-chain geometries are within the acceptable limits of published values[12]. The location of the zinc-binding site in the structure of domain I, combined with previous biochemical characterization, suggests that zinc has a structural role in the maintenance of the NS5A fold. Following on from the zinc-coordination site, a stretch of random coil exits H2 and passes across the β-sheet and N-terminal strand, in an orientation orthogonal to the plane of the sheet. This coiled sequence then enters a tight turn and runs adjacent to the β-sheet, away from the zinc atom-coordination site, before entering a proline-rich region that connects subdomain IA to subdomain IB.

Subdomain IB consists of a four-strand anti-parallel β-sheet (B4–B7) and a small, two-strand anti-parallel β-sheet (B8 and B9) near the C terminus, surrounded by extensive random coil structures (best seen in Fig. 1d). Perhaps the most surprising observation from model building and refinement of the structure was the presence of a disulphide bond near the C terminus of domain I. The disulphide bond connects the side chains of the conserved residues Cys 142 and Cys 190, resulting in a covalent link between the loop exiting from β-strand B6 to the C-terminal extension of strand B9. Model refinement without the disulphide bond placed the side chains of these cysteine residues in an unfavourable proximity. Refinement with the disulphide bond led to no problematic geometry for either cysteine, and generated a model that better fitted the electron density (Fig. 2c). Density corresponding to the disulphide bond is present in both molecules in the asymmetric unit, providing two independent views of this feature. The disulphide bond in the model has in a sulphur-to-sulphur atom distance of 2.03 Å, an ideal value for bond formation[13]. Although all evidence points to the presence of a disulphide bond in the crystal, it is not yet clear whether this bond exists in NS5A in the context of an HCV infection. Mutagenesis of these cysteine residues (Cys 142 and Cys 190) produced only a mild defect in RNA replication, indicating that the disulphide bond is not required for the RNA replicase functions of NS5A[10]. However, it is enticing to imagine that the disulphide bond has a regulatory role in NS5A function: by tethering the C terminus of domain I and probably altering the arrangement of C-terminal domains II and III, it might serve as a conformational switch to modulate replicase versus non-replicase activities of NS5A.

Analysis of the molecular surface of domain I is important to furthering our understanding of NS5A. Figure 3a presents two rotational views of the solvent-accessible molecular surface of NS5A domain I, coloured according to electrostatic potential. The orientations (left to right) are identical to those shown in the ribbon diagrams in Figs 1b and c, respectively. These views highlight the uneven charge distribution within domain I. The N-terminal subdomain IA region has an almost exclusively basic surface, whereas the C-terminal subdomain IB is predominantly acidic. This unusual charge distribution has implications for the dimeric form of the protein.

We analysed the conservation of surface-exposed residues of domain I in order to define potential interaction sites. A conservation plot, generated on the basis of sequence alignments of domain I

[1]Laboratory of Virology and Infectious Disease, Center for the Study of Hepatitis C, The Rockefeller University, 1230 York Avenue Box 64, New York, New York 10021, USA.

© 2005 Nature Publishing Group

regions from the 30 HCV genotype reference sequences in the Los Alamos HCV database[14] (http://gluttony.lanl.gov/content/hcv-db/), as well as the related hepatitis GB virus B, is shown in Fig. 3b. The plot shows the substantial overall surface conservation of domain I, and highlights a large patch of conserved residues that might represent a molecular interaction surface. This contiguous surface spans subdomain IA and IB, including a 'pocket' generated at the interface between these elements. A plot of electrostatic potential of this same surface is shown, highlighting the complex mixture of charged and hydrophobic residues generating this surface (Fig. 3c). A ribbon diagram of this orientation of domain I is also shown (Fig. 3d). A large number of proteins have been shown to interact with domain I (ref. 6). Whether a cellular protein or a viral component of the replicase (including the other domains of NS5A) interacts with this surface remains to be determined.

Another surface of interest was the region buried between the monomers to create the dimeric domain I seen in the crystal structure. Figure 3e presents a surface conservation plot highlighting these interaction patches (orientation similar to Fig. 1d). The dimer interface (678 $Å^2$) consists of two patches of buried surface area, one located primarily in subdomain IA, and a second located in subdomain IB. The total buried surface area in the domain I dimer is larger than the generally accepted standard of 600 $Å^2$ for protein interfaces, and has good shape and electrostatic complementarity[15]. The contact patch in subdomain IA contains a number of conserved residues, including several residues surrounding the zinc-coordination site of domain I. The smaller patch in subdomain IB has lower sequence conservation. It should be noted that residues of lower conservation at in these interfaces appear to be involved primarily in main-chain contacts.

Homotypic oligomeric interactions of the NS5A protein have been described[16]. Analytical ultracentrifugation results indicate that domain I is monomeric, although it should be noted that these experiments were performed with only a portion of the NS5A protein, in the absence of membranes and RNA. It is likely that in the context of the HCV replicase complex, NS5A is present at very high local concentrations that might drive oligomerization. In this respect, NS5A might be similar to the NS5B protein, which is monomeric in solution but readily forms 2-dimensional arrays on membrane surfaces[17]. Perhaps the conditions that favour protein–protein interactions in crystallization also mimic some of the relevant NS5A oligomerization interactions. Although more work is needed to determine the oligomeric state of NS5A in the replicase, the dimeric form presented here identifies a number of features that might have relevance to HCV biology. Three rotations of the dimer are presented in Fig. 4a, showing the interface between

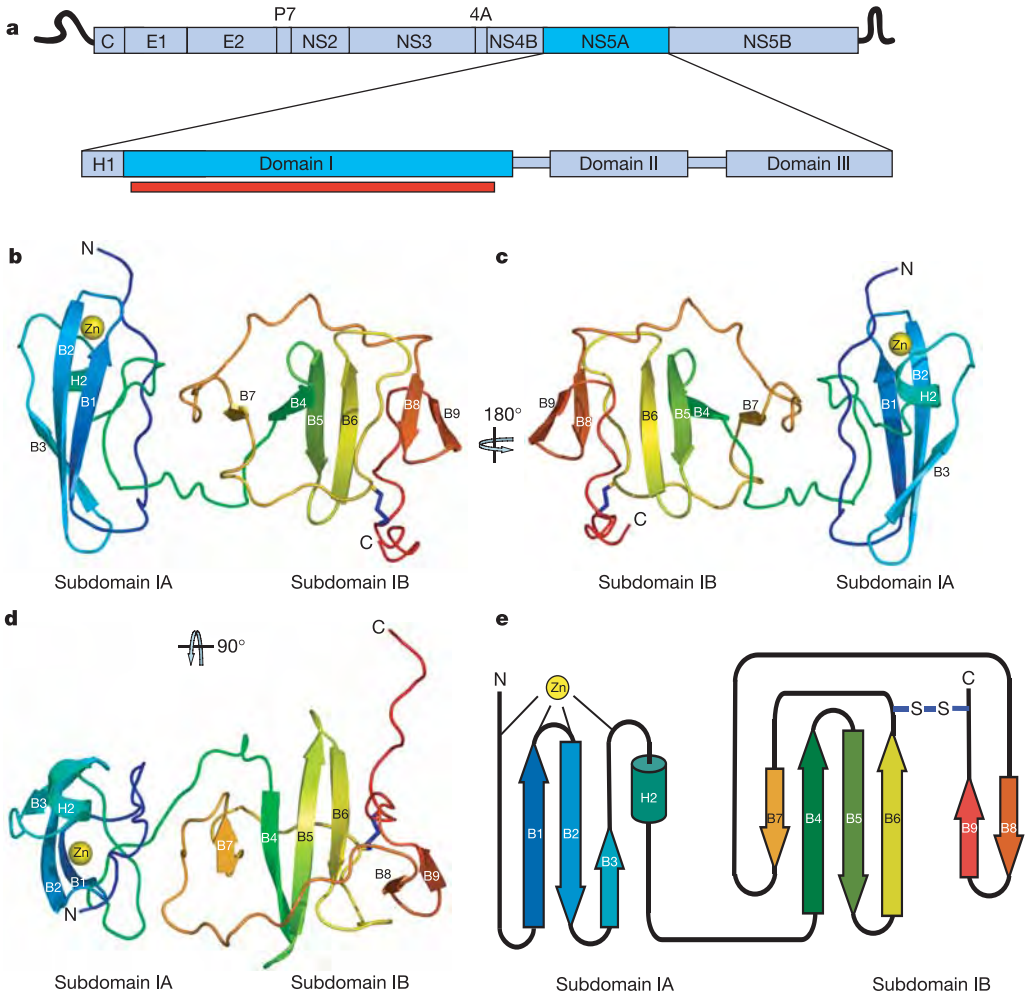


Figure 1 | An overview of the NS5A domain I structure. a, Schematic of HCV genome organization and the domain structure of the NS5A protein. The portion of domain I in presented in this crystal structure is indicated by the red bar. C, capsid protein; E1 and E2, envelope glycoproteins 1 and 2 (respectively); 4A, NS4A protein. **b**, Ribbon diagram of the structure of domain I. The polypeptide chain is coloured from the N terminus (blue) to C terminus (red). The coordinated zinc atom is shown in yellow. The C-terminal disulphide bond is shown in blue. **c**, A 180° rotation showing the 'back' of domain I. **d**, A 90° rotation showing the 'top-down' view of domain I. **e**, Domain I topology organization model.

© 2005 Nature Publishing Group

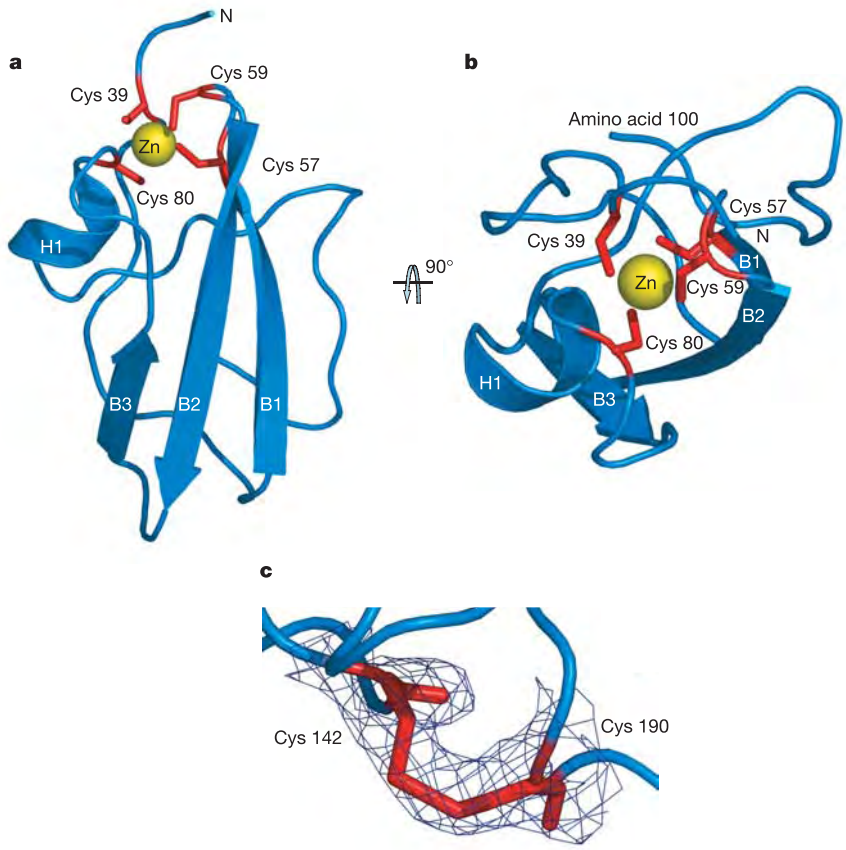


Figure 2 | The NS5A zinc-binding motif and disulphide bond. **a**, A view of subdomain 1A (amino acids 36–100), highlighting the zinc-binding motif. The zinc atom (yellow) is coordinated by four cysteine residues (red). **b**, A 'top-down' view of the zinc-coordination site, showing caging of the zinc atom. **c**, Overlay of the experimental electron-density map on the model of the disulphide bond at 1σ. Amino acid residues Cys 142 and Cys 190 are labelled.

Table 1 | Summary of crystal parameters, data collection and refinement statistics

Crystal parameters Spacegroup	Unit cell lengths (a,b) (Å)	Unit cell lengths (c) (Å)	Unit cell angles (α,β,γ)			Molecules per asymmetric unit
$P4_122$	55.28	312.30	90°			2
Data collection Wavelength (Å)	Resolution (Å)	Reflections measured/unique	Completeness* (%)	R_{sym}*† (%)	$I/\sigma(I)$*	Phasing power‡ (anomalous)
$\lambda1 = 1.28345$	30.0–2.50	301,420/17,992	100 (99.9)	6.9 (32)	23.8 (4.7)	1.18
$\lambda2 = 1.28385$	30.0–2.50	302,219/17,992	100 (99.9)	4.5 (34)	22.4 (4.3)	1.25
$\lambda3 = 1.27069$	30.0–2.50	314,341/18,007	100 (100)	8.7 (27)	35.6 (5.7)	1.21
Figure of merit	Acentric 0.48	Centric 0.40	Overall 0.46			
Refinement against λ3 Resolution (Å)	Cut-off	Reflections	Completeness (%)	R_{cryst}§ (%)		R_{free}‖ (%)
30.0-2.50	$\lvert F\rvert/\sigma\lvert F\rvert > 2.0$	29,570	92.5	22.5		28.5
Root mean square deviations Molecules Cα (Å)	Bond lengths (Å)	Bond angles	Thermal parameters main chain atoms (Å^2)			Thermal parameters side chain atoms(Å^2)
0.43	0.00633	1.35°	1.12			1.80

*Values reported in the format: overall data (last resolution shell).
†$R_{sym} = \Sigma|I - \langle I \rangle|/\Sigma I$, where I is the observed intensity and $\langle I \rangle$ is the average intensity obtained from multiple observations of symmetry-related reflections.
‡Phasing power = r.m.s. ($|F_H|/E$), where $|F_H|$ is the heavy atom structure factor amplitude and E is the residual lack of closure.
§$R_{cryst} = \Sigma|F_{obs} - F_{calc}|/\Sigma|F_{obs}|$, where F_{obs} and F_{calc} are the observed and calculated structure factors, respectively.
‖R_{free} is the same as R_{cryst} but is calculated with 10% of the data.

© 2005 Nature Publishing Group

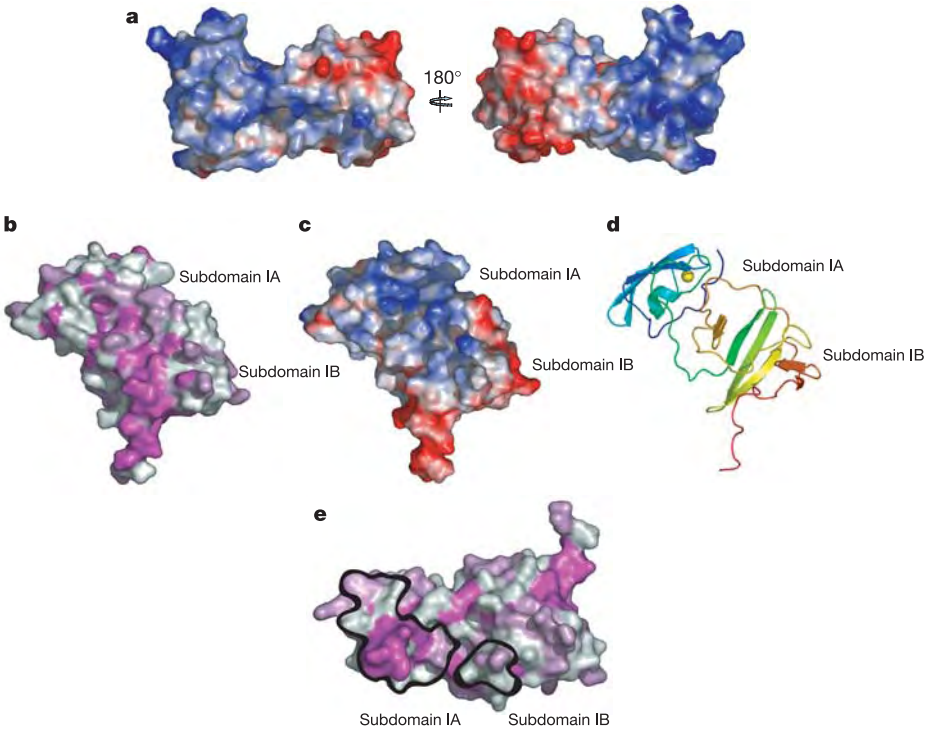


Figure 3 | Molecular surfaces of domain I. a, Domain I surface potential, with acidic surfaces in red, neutral surfaces in white and basic surfaces in blue. Two rotations of domain I are shown, corresponding to the views in Fig. 1b and c, respectively. **b**, Surface of domain I coloured according to residue conservation: magenta, ≥95% conservation; light pink, 75–95% conservation; white, ≤75% conservation. Domain I is oriented to show the most conserved molecular surface. **c**, A surface-potential view of the image in **b**. **d**, A ribbon diagram view of the orientation in **b** and **c**. **e**, Conservation plot showing the dimer interface surfaces (outlined in black).

monomers and the large groove between the domain IB regions that is generated by the 'claw-like' shape of the dimer.

As NS5A probably makes contacts with proteins, RNA and membranes, an analysis of electrostatic surface potential is important for determining how a dimeric domain I might fit into the replicase (Fig. 4b). One notable feature in these surface-potential plots is the basic surface of domain I near the N termini of the dimer. An NMR structure of the membrane anchor of NS5A has recently been determined[8]. This anchoring helix is five residues from the N terminus of the domain I structure, suggesting that domain I is close to the membrane. The basic nature of the surface of domain I close to the anchor helix is consistent with this, as the protein is probably in close contact with the negatively charged head groups of the membrane. This interaction would position the domain I dimer groove facing away from the membrane, where it could interact with RNA. Interactions of NS5A with RNA during protein purification have been described[18], although specific binding remains to be demonstrated.

This identified groove is an attractive RNA-binding pocket, especially when surface potentials of the groove are plotted. Modelling suggests the groove is of sufficient dimensions to bind to either single- or double-stranded RNA. The deep, highly basic portion of the groove has a diameter of 13.5 Å. The groove then expands to diameter of approximately 27.5 Å in a less charged, polar boundary region. An RNA molecule could easily fit in the groove, making both electrostatic contacts with the deep basic groove and other contacts with the groove boundary region. An exposed tryptophan residue (Trp 84) lies on the floor of the basic groove, where it could interact with RNA through stacking interactions. The protein 'arms' extending out past this groove are more acidic, and might serve to prevent RNA from exiting the groove. The positioning of NS5A domains II and III relative to the groove is unclear, but we might imagine these domains to interact with RNA positioned by the domain I groove. Furthermore, the dimeric form of domain I places the conserved interaction surface described for the monomer on the outside of the 'arms' of the dimer, where it would be available for interactions. A model of the NS5A dimer in relation to the cell membrane and the N-terminal membrane anchor is presented in Fig. 4c. The sequence of domain I, highlighting secondary structures and important residues and surfaces, is provided in the Supplementary Information.

The crystal structure of the domain I region of NS5A provides the first molecular view of this important protein, and might reflect a new class of viral replicase proteins. The high-resolution view of domain I has practical applications for antiviral drug design, and provides a rational framework for experimentally addressing NS5A function in the HCV replicase and in the regulation of cellular processes. These experiments might ultimately provide clues as to how the replicase machinery works to replicate HCV RNA.

METHODS

Protein preparation. The expression and purification of genotype 1b, Con1 subtype NS5A Domain I-Δh (containing a deletion of the membrane anchor helix) was performed as described previously[10]. Following initial purification, the C-terminal polyhistidine tag and linker sequence were proteolytically removed from NS5A-Domain I-Δh by overnight incubation at 25 °C in the presence of recombinant light-chain enterokinase protease. Once cleavage was complete, the protein was re-purified by ion exchange on a HiPrep 16/20 DEAE column (Amersham/Pharmacia) using previously described conditions[10] to separate the protease and tag from the Domain I-Δh protein. The protein was then exchanged into buffer A (25 mM Tris-HCl pH 8.0, 175 mM NaCl and 20% (v/v) glycerol) using a HiPrep 26/10 desalting column (Amersham/Pharmacia). Protein was then concentrated to 20–30 mg ml^{-1} using Amicon ultra 4 centrifugal concentrators (Millipore). Protein yields were typically 10–12 mg l^{-1} of bacterial culture at an estimated 95% purity.

© 2005 Nature Publishing Group

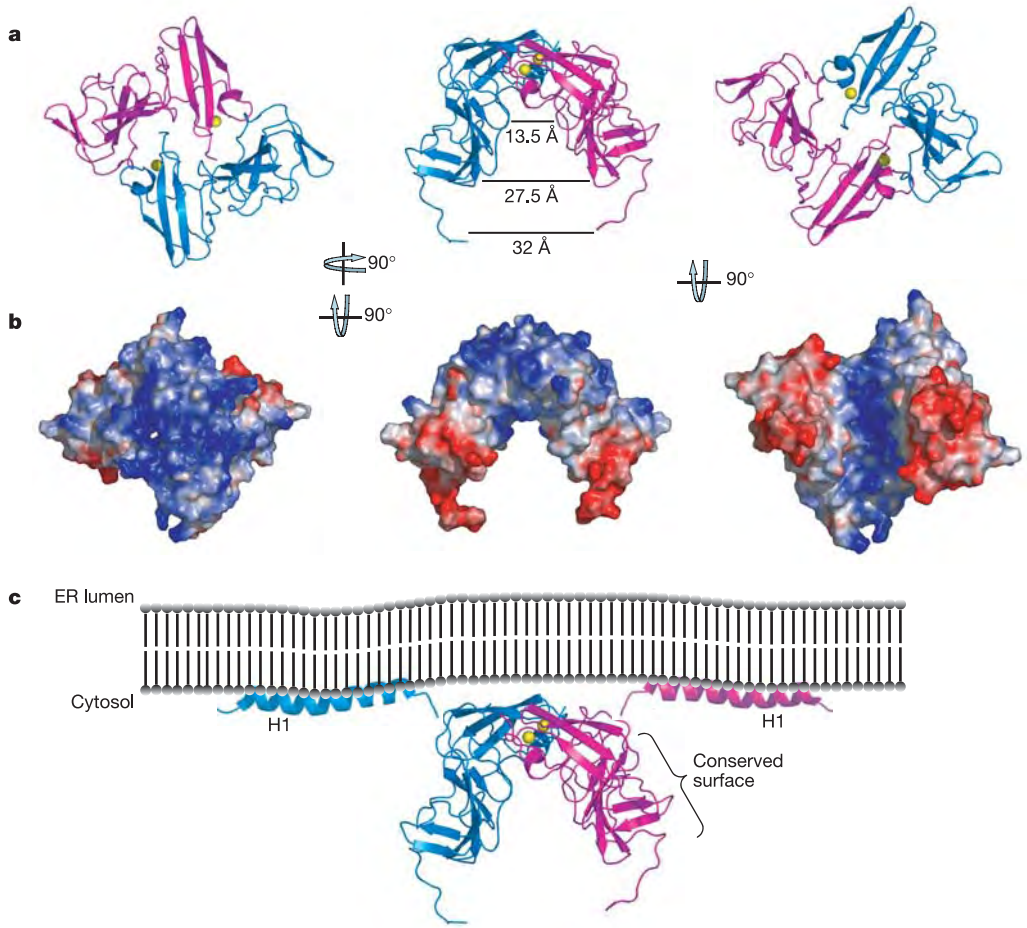


Figure 4 | The NS5A domain I dimer reveals potential interaction surfaces. a, Ribbon diagrams of three rotations of the domain I dimer. **b**, Surface-potential plots of the domain I dimer, with views corresponding to those shown in **a**. Analysis of images in **a** and **b** shows that the dimer creates a relatively flat, basic surface near the N terminus and a large groove between the two subdomain IB regions. **c**, Model of NS5A position relative to the endoplasmic reticulum membrane. The location of the conserved surface in Fig. 3b is indicated. Helices are from the recent structure of the NS5A N-terminal helix[8].

Crystal growth and freezing. Crystals of NS5A Domain I-Δh were grown by hanging-drop vapour diffusion at 4 °C on siliconized coverslips in 24-well Linbro plates. The 1 ml of well solution contained 100 mM HEPES pH 7.0, 0.6 M trisodium citrate dihydrate and 13% (v/v) isopropanol. The drop contained 1.5 μl of Domain I-Δh protein at 26 mg ml^{-1} in buffer A, 1.25 μl well solution and 0.5 μl of 2 M non-detergent sulphobetaine 201. Crystals of 0.15–0.3-mm size grew from these conditions in approximately eight days. For freezing, crystals were transferred from hanging drops to a 2-μl drop of well solution plus 20% (v/v) glycerol, and incubated at 4 °C for 10 min. Crystals were then transferred to a 2-μl drop of buffer A with a total of 30% (v/v) glycerol and allowed to equilibrate for 10 min. Crystals were then harvested and flash frozen in liquid propane.

Data collection. All data collection was performed at beam line X9A at the Brookhaven National Labs National Synchrotron Light Source. Phases were determined from the endogenous zinc present in NS5A. Before diffraction data collection, X-ray fluorescence scans were used to confirm the presence of zinc in the protein crystals and determine the K absorption edge of zinc. For zinc multiwavelength anomalous diffraction (MAD), data were collected at the zinc absorption maxima (9,660 eV, 1.28345 Å), the inflection point maxima (9,657 eV, 1.28385 Å) and at a remote wavelength (9,757 eV, 1.27069 Å). Data collection was limited to 2.5-Å resolution by a combination of detector geometry and the presence of an unusually long axis in the crystal.

Data processing and model building. Data were processed and scaled using DENZO/SCALEPACK[19]. The data for each of the three MAD wavelength data sets were then used to locate anomalous peaks, corresponding to the presence of the zinc atom in NS5A. The two zinc atom sites were found using the program SOLVE[20]. An interpretable electron-density map was obtained using MLPHARE, followed by density modification and phase combination using DM[21]. The inflection data set was used for electron-density map calculation. Several rounds of iterative model building and refinement were performed using the programs O[22] and CNS[23]. The final model contained 96 solvent molecules, two zinc atoms and two molecules of NS5A (consisting of amino acids 36–198 of NS5A, with density not observed for the extreme N and C termini). A summary of the final refinement statistics is provided in Table 1. PROCHECK[24] revealed no unfavourable (ϕ,ψ) combinations, and main-chain and side-chain structural parameters were consistently better than or within the average for structures refined to 2.5 Å. Atomic coordinates have been deposited in the Protein Data Bank under accession number 1ZH1. Secondary structures were assigned using the program DSSP[25]. Graphics presented in this manuscript were generated using the program PyMOL[26]. APBS was used for calculating surface potentials[27]. Sequence alignments were performed using ClustalX[28], and plotting of conservation to molecular surfaces was performed using msf_similarity_to_pdb (David Jeruzalmi, personal communication).

Received 24 January; accepted 1 April 2005.

1. World Health Organization. Hepatitis C: global prevalence. *Wkly. Epidemiol. Rec.* **72**, 341–344 (1997).
2. Blight, K. J., Kolykhalov, A. A. & Rice, C. M. Efficient initiation of HCV RNA replication in cell culture. *Science* **290**, 1972–1974 (2000).
3. Lohmann, V. *et al.* Replication of subgenomic hepatitis C virus RNAs in a hepatoma cell line. *Science* **285**, 110–113 (1999).
4. Lohmann, V., Korner, F., Dobierzewska, A. & Bartenschlager, R. Mutations in hepatitis C virus RNAs conferring cell culture adaptation. *J. Virol.* **75**, 1437–1449 (2001).
5. Tellinghuisen, T. L. & Rice, C. M. Interaction between hepatitis C virus proteins and host cell factors. *Curr. Opin. Microbiol.* **5**, 419–427 (2002).

© 2005 Nature Publishing Group

6. Macdonald, A. & Harris, M. Hepatitis C virus NS5A: tales of a promiscuous protein. *J. Gen. Virol.* **85,** 2485–2502 (2004).
7. Elazar, M. *et al.* Amphipathic helix-dependent localization of NS5A mediates hepatitis C virus RNA replication. *J. Virol.* **77,** 6055–6061 (2003).
8. Penin, F. *et al.* Structure and function of the membrane anchor domain of hepatitis C virus nonstructural protein 5A. *J. Biol. Chem.* **279,** 40835–40843 (2004).
9. Brass, V. *et al.* An amino-terminal amphipathic α-helix mediates membrane association of the hepatitis C virus nonstructural protein 5A. *J. Biol. Chem.* **277,** 8130–8139 (2002).
10. Tellinghuisen, T. L., Marcotrigiano, J., Gorbalenya, A. E. & Rice, C. M. The NS5A protein of hepatitis C virus is a zinc metalloprotein. *J. Biol. Chem.* **279,** 48576–48587 (2004).
11. Holm, L. & Sander, C. Mapping the protein universe. *Science* **273,** 595–603 (1996).
12. Alberts, I. L., Nadassy, K. & Wodak, S. J. Analysis of zinc binding sites in protein crystal structures. *Protein Sci.* **7,** 1700–1716 (1998).
13. Engh, R. & Huber, R. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr. A* **47,** 392–400 (1991).
14. Kuiken, C., Yusim, K., Boykin, L. & Richardson, R. The Los Alamos HCV sequence database. *Bioinformatics* **21,** 379–384 (2005).
15. Jones, S. & Thornton, J. M. Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA* **93,** 13–20 (1996).
16. Dimitrova, M., Imbert, I., Kieny, M. P. & Schuster, C. Protein–protein interactions between hepatitis C virus nonstructural proteins. *J. Virol.* **77,** 5401–5414 (2003).
17. Wang, Q. M. *et al.* Oligomerization and cooperative RNA synthesis activity of hepatitis C virus RNA-dependent RNA polymerase. *J. Virol.* **76,** 3865–3872 (2002).
18. Huang, L. *et al.* Purification and characterization of hepatitis C virus non-structural protein 5A expressed in *Escherichia coli*. *Protein Expr. Purif.* **37,** 144–153 (2004).
19. Otwinowski, Z. & Minor, M. in *Methods in Enzymology Vol. 276, Macromolecular Crystallography, Part A* (eds Carter, C. W. & Sweet, R. M.) 307–326 (Academic, New York, 1997).
20. Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **55,** 849–861 (1999).
21. Collaborative Computational Project, N. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **50,** 760–763 (1994).
22. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).
23. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54,** 905–921 (1998).
24. Laskowski, R. A., Moss, D. S. & Thornton, J. M. Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.* **231,** 1049–1067 (1993).
25. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–2637 (1983).
26. DeLano, W. L. The PyMOL molecular graphics system. ⟨http://www.pymol.org⟩ (2002).
27. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl Acad. Sci. USA* **98,** 10037–10041 (2001).
28. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **24,** 4876–4882 (1997).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to acknowledge R. MacKinnon, S. Darst and H. Mueller for the use of X-ray diffractometers, related equipment and software. We appreciate access to beamline X9A at the National Synchrotron Light Source (NSLS) at the Brookhaven National Labs and acknowledge the assistance of NSLS staff. J.-W. Carroll provided vital assistance with data collection. D. Jeruzalmi provided the program msf_similarity_to_pdb. We wish to thank S. Darst, M. Evans, A. Gauthier, C. Jones, B. Lindenbach, I. Lorenz, T. von Hahn, M. Tellinghuisen and K. Tellinghuisen for critical reading of this manuscript. T.L.T. was supported in part by fellowships from the Charles H. Revson Foundation for Biomedical Research and the National Institutes of Health Ruth L. Kirschstein National Research Service Award, granted through the National Institute of Allergy and Infectious Disease. J.M. was supported as a Merck Fellow of the Life Sciences Research Foundation. Additional financial support for this work came from grants from the National Institutes of Health and the Greenberg Medical Research Institute (C.M.R.).

Author Contributions T.L.T., J.M. and C.M.R. conceived these experiments. T.L.T. generated all reagents, materials, proteins and crystals used herein, with assistance from J.M. All data collection was carried out by J.M. and T.L.T. Data processing, model building and refinement were performed by T.L.T., with significant input and assistance from J.M. The manuscript was written by T.L.T. with comments and assistance from J.M. and C.M.R.

Author Information The coordinates for this structure have been deposited in the Protein Data Bank (PDB) under accession code 1ZH1. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.M.R. (ricec@rockefeller.edu) or J.M. (marcotj@rockefeller.edu).

© 2005 Nature Publishing Group

THE CAREERS MAGAZINE FOR SCIENTISTS

Vol 435 | Issue no. 7040 | 19 May 2005

FOCUS
SPOTLIGHT
RECRUITMENT
ANNOUNCEMENTS
EVENTS

naturejobs

To plan...or not to plan?

Many senior scientists achieved success without drawing up a clear path for their career beforehand — a fact emphasized by several speakers at a recent careers fair. But those same panellists also affirmed the need for planning. This apparent paradox confused many who attended the New York fair, which was sponsored by *Naturejobs* and the New York Academy of Sciences. Indeed, many young scientists returned to the topic time and again during the subsequent question-and-answer sessions.

Keynote speaker Peter Corr set both the theme and the tone of the questions that would follow. Senior vice-president for science and technology at drugs firm Pfizer, Corr talked about how, as a young academic climbing the tenure track, he couldn't have imagined that he would one day be a manager at the world's biggest pharmaceutical company.

Later panellists described how they too had had difficulty predicting their future. But they said you should be prepared to try — especially in job interviews, which can present a tricky intersection between selling your ambitions either too long or too short. The trick is to show that you want to grow and learn, but not necessarily give too much detail, said Grace Wong, chief scientific officer of ActoKine Therapeutics, a biotech company in Chestnut Hill, Massachusetts.

Of course, there are exceptions, said Scott Wadsworth, a research fellow at pharmaceutical company Johnson & Johnson. For example, a candidate who plans to work in a company for a few years and then go on for medical training would be welcome at a drugs firm because their desire to learn would still benefit the company, Wadsworth said.

Corr's opening remarks actually supplied some answers to the questions that arose in later sessions. Rather than trying to predict what the next job or position is, he said, you should learn something new every day, work on projects outside your skill set and comfort zone, and treat people well. The future should then take care of itself.

Paul Smaglik, *Naturejobs* editor

CONTACTS
Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

US Head Office, New York
345 Park Avenue South, 10th Floor,
New York, NY 10010-1707
Tel: +1 800 989 7718
Fax: +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager/Corporations:
Peter Bless
Classified Sales Representatives
Tel: +1 800 989 7718

New York/Pennsylvania/
Latin America: Kelly Roman
Midwest USA/Maryland/
NIH: Wade Tucker
East USA/Canada:
Janine Taormina

San Francisco Office
Classified Sales Representative:
Michaela Bjorkman
West USA/West Corp. Canada
225 Bush Street, Suite 1453
San Francisco, CA 94104
Tel: +1 415 781 3803
Fax: +1 415 781 3805
e-mail: m.bjorkman@naturesf.com

European Head Office, London
The Macmillan Building,
4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0) 20 7843 4961
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

***Naturejobs* Sales Director:** Nevin Bayoumi (4978)
European Sales Manager: Andy Douglas (4975)

Advertising Production Manager: Billie Franklin
To send materials use London address above.
Tel: +44 (0) 20 7843 4814
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

***Naturejobs* web development:** Tom Hancock
***Naturejobs* online production:** Niamh Shields

European Satellite Office
Germany/Austria/Italy/
The Netherlands/Belgium:
Patrick Phelan, e-mail: p.phelan@nature.com
Reya Silao. e-mail: r.silao@nature.com

Japan Head Office, Tokyo
MG Ichigaya Building (5F), 19-1 Haraikatamachi,
Shinjuku-ku, Tokyo 162-0841
Tel: +81 3 3267 8751
Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
e-mail: r.asami@naturejpn.com

©2005 Nature Publishing Group

NATURE|Vol 435|19 May 2005

MOVERS

Robert Rosner, laboratory director, Argonne National Laboratory, Illinois

2002–05: Chief scientist, Argonne National Laboratory
1997–02: Director, Center for Astrophysical Thermonuclear Flashes at Chicago
1991–97: Chairman, Astronomy and Astrophysics, University of Chicago, Chicago, Illinois

Staying true to Robert Rosner's two great passions — his wife and scientific discovery — set him on an astronomical career path. To avoid a commuter marriage, Rosner decided to stay in the Cambridge area after finishing his physics PhD at Harvard University, while his wife completed her doctoral degree — even if it meant changing fields. He became, in effect, the 'accidental' astrophysicist.

Rosner was offered a postdoc with Giuseppe Vaiana, who had just moved to Harvard. Vaiana was a leader in experimental solar physics, and one of the first people to use imaging X-ray telescopes to look at the Sun. "He, more than anyone else, channelled my passion for astrophysics," says Rosner.

Now that he knows an unplanned foray can dramatically influence your job path, Rosner advises young scientists to anticipate that even the best-laid career plans are likely to change. "I hadn't realized to what extent the postdoc would mould the rest of my career. I thought it was an interlude," he says.

After 17 years in the Boston area and a tenured position with the Smithsonian Astrophysical Observatory, Rosner left Boston with his wife. They both took positions at the University of Chicago.

In the early 1990s, Rosner was asked by Chicago's provost to head a committee to examine the relationship between the university and Argonne National Laboratory, which the university operates. The time spent thinking about the role that national labs have to play within national research agendas awakened Rosner's desire for furthering US excellence in science and technology. Three years ago, he joined Argonne on a part-time basis as chief scientist.

He says that his new role as lab director is the greatest challenge of his career, particularly because scientists rarely get formal training in leadership and management. But he says that his on-the-job experience with past mentors, including his predecessor at Argonne, has prepared him for the task at hand.

One of his first tasks, he says, will be to secure the development of a new accelerator — the rare isotope accelerator — to open fresh frontiers in nuclear physics. And he has a broader goal: to make sure that Argonne seizes every opportunity it can to play a key role in maintaining the United States' leading position in science and technology. ■

RECRUITERS & ACADEMIA

Zen and the art of grant application

The grant-writing seminars I took did not fully prepare me for the actual process of applying for a grant from the US National Institutes of Health (NIH). First was the time issue. I applied for a grant to study a protein found in relative abundance in a cell's endoplasmic reticulum. This protein has no known function, and I was determined to find it.

I sent the grant application in June and didn't get summary statements back until December, which meant that, with luck, I could address the issues raised and resubmit by June. I had to take a deep breath and swallow the realization that it would take an entire year just to resubmit my application. For my next submission, I will plan for this long period — while hoping for a faster turn-around.

I also became trapped in a catch-22. It is very difficult to get funding for something, such as my protein, that is not well understood. The thinking, perhaps, is that there are so many things to study that we do understand, why add things we know nothing about? This leaves me scratching my head as to how anyone will ever get funding to study the large number of proteins of 'unknown function' that have been identified by the Human Genome Project.

To overcome this apparent paradox, it seems you need to find something that is understood or that you can flesh out, or you must search the literature for something that seems to explain to some extent what you propose to study. I found a paper showing that my protein could bind and activate tissue plasminogen activator, an important molecule for the regulation of fibrinolysis in the vascular system and something related to human health (which is a plus when seeking funds).

I discovered from reviewers' comments that it pays to be as detailed as possible, and so build their confidence in your ability to carry out the experiment. I learned this the hard way: a reviewer said my description of the techniques and experimental approaches were too brief and superficial — and that I did not have an established track record with the studies. I had no doubt about my ability to do the experiments, and I had included in the initial application information that I had used the technique before. But I realized that writing down exact protocols might provide that added information to boost reviewer confidence in my abilities.

All told, deep breaths and an eye to including even what you may consider obvious should help you to steer your application to success. ■

Dorothy Mundy is an assistant professor in the Department of Cell Biology, University of Texas Southwestern Medical School, Dallas.

NIH Office of Extramural Research, Grants
grants1.nih.gov/grants/oer.htm

GRADUATE JOURNAL

Dissin' the dissertation

A graduate career is bookended by the Graduate Record Examination and the dissertation, and to me, neither seems all that useful. Allow me a moment to rant on about why scientists should not have to write dissertations. Am I the only one who thinks this is an antiquated system?

Sure, the history and literature students write a big book to form the bulk of their PhDs, but do any potential labs or colleges ever ask to see a scientist's dissertation upon application? They open up the manila envelope (or e-mail these days) and take a quick glance at the CV with their eyes shooting straight to the publication list. Our graduate careers are defined by what we learn and what articles we get into which journals.

So, here I sit, bitter and with dissertation.doc as a constant open window on my laptop. I peck away at it and thank my lucky stars that I will never have to do this again. I have jumped through so many hoops during the graduate-school circus and this one just seems so silly. Will anybody ever read this book I'm writing? Unlikely. But, for now, I'm back to working on it.

It's going to be a crazy two weeks ahead. Send pizza, coffee and beer... ■

Jason Underwood is a graduate student in molecular biology at the University of California, Los Angeles.

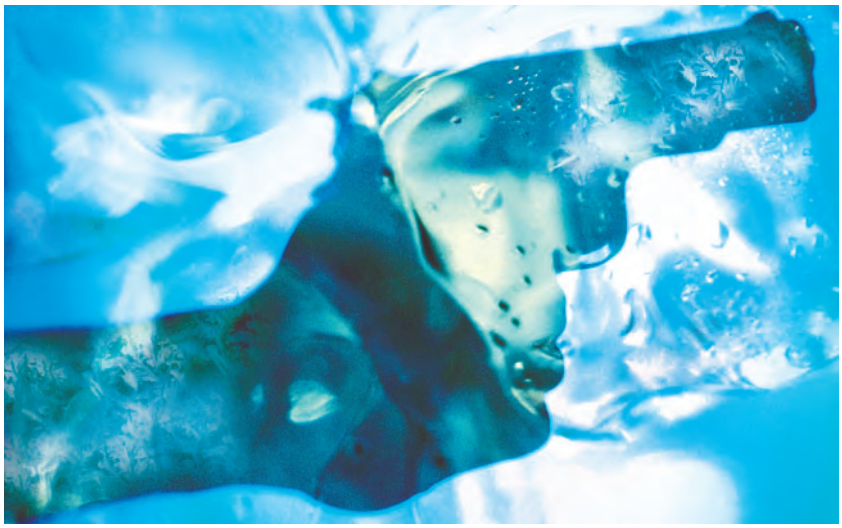
©2005 Nature Publishing Group

NATURE|Vol 435|19 May 2005

Climate change

What will happen in the cold light of day?

Michael Hanlon



JACEY

"Please, in the name of all the world, not just Russia, you must sign this. It is our only hope." Nadezhda Ledenaya, Politburo member, environment minister and the USSR's chief negotiator was close to tears as she realized the 2006 Reykjavik Protocol was dead in the water. Outside, the birds sang in the warm April sunshine, drowned out by the chants of the eco-warriors and the cackle of reporters on their cellphones.

The third meeting of the Intergovernmental Panel on Climate Change had always promised to be a brutal affair. There were the usual suspects on the fringes, the non-governmental organizations and green advocates arguing that the Protocol did not go far enough. At the centre of things were the major governments, which seemed to be on the verge of agreement until that fatal, last-minute deal, cooked up over a secret power-breakfast at the Borg. The deal united the Japanese, Chinese and Europeans, and Reykjavik became meaningless, even though it had the full backing of the US–Canadian–Soviet alliance. Of course, the Protocol went through — but without Europe and East Asia on board, how much difference would it really make?

The world first began to take notice of the threat posed by global cooling in the late 1980s. Fusion power, which had swept all before it after that rather unexpected breakthrough in 1969, had seemed at the time to be the answer to all of humanity's problems. For a start, it meant the end of the old fission stations, which were becoming an embarrassment after the terrible accidents of '63. And the switch away from fossil fuels meant that half a century of squabbling and war over oil was ended.

But the fusion reactors created their own problems. The huge seawater processing plants were responsible for the production of millions of tonnes of microscopic salt particles, all thrown into the atmosphere. This, of course, was what was being blamed for the biggest scare-story since the end of the Wars — catastrophic climate change.

At the moment the doom brigade had the upper hand. The globe had already cooled by some 1.3 °C in the 38 years since fusion had been cracked. There was debate over this — measurements in the late 1990s by Soviet and US satellites seemed initially to contradict the ground data, but now even the sceptics agreed that something was happening. The only real debate was over how much of this cooling was caused by the extra cloud cover generated by the salt aerosols, and how much could put be down to natural causes.

Debate was fiercer about the future. At the extreme end, the hyper-coolers predicted a 10 °C chilling by 2100. This, everyone agreed, would be a true cataclysm. Even the 'official' estimate of a four-degree drop would lead to a massive alteration in the global climate.

There was talk of losing the Siberian and Canadian grain and cattle belts, hence the near-panic among the Soviet delegates at Reykjavik. What was now fertile farmland would be turned into unproductive tundra in decades. The English, Danish and Irish wine industries would be decimated. The Neocom Soviet Union, whose vast wealth was predicated upon its status as the bread-basket of the world, would be thrown into penury and, potentially, millions could starve.

And although some wags speculated that the lush, heathered hills of Scotland could one day be alive to the braying of skiers and snowboarders, others worried that glaciers could reform in the Alps, for the first time in 180,000 years, threatening dozens of towns and villages.

There were even lurid stories, written by scores of environmental reporters, that if global cooling was not checked, the great Arctic seaways could be blocked by ice.

Of course not everyone agreed with these alarming predictions. Sceptics pointed to the fact that the recent 'cliff' seen in the climate graph — a sickening drop of nearly half a degree since 1990 — could be a statistical anomaly. The fact that many of these sceptics seemed to be associated with some of the more dubious fusion enterprises did not help their case, but they made some compelling arguments.

Probably the most powerful were made by the Swedish gadfly Anna Carlsson, who pointed out that phasing out the fusion stations and switching to coal, oil and gas or even wind, solar and wave power would probably cripple the world's economy. But her (more subtle) argument was that there was no real evidence that a cooler world would be a disaster. For long periods of our planet's history, after all, both poles had been covered by more-or-less permanent ice, and sea levels today are at almost unprecedented high levels.

As the dust settled on the Reykjavik documents, world leaders scurried back to the airport in their whispering electric limos. Fusion power had cemented and underpinned the great alliance, the Soviet–American economic and military pacts of the '70s and '80s, which meant it now seemed unthinkable that these great powers could ever go to war.

But it now looked possible — likely, even — that we could see a return to the bad old days. If the cooling really made itself felt, then from under the sleepy forests of Araby would flow the solution. With no deal, the world would be forced into action — and as far as energy goes it would be a seller's market.

We would all be looking down the barrel of a gun and there would be absolutely nothing we could do. ■

Michael Hanlon is the science editor of the *Daily Mail*. His new book, *The Science of the Hitchhiker's Guide to the Galaxy*, is published by Macmillan (for review see *Nature* **435,** 148; 2005).

FUTURES

©2005 Nature Publishing Group